Network-Based Methods for the Analysis of Next Generation Sequencing Data in Human Genetic Disease

Dand, Nicholas James

Awarding institution:
King's College London

Network-Based Methods for the Analysis of Next Generation Sequencing Data in Human Genetic Disease

Nicholas James Dand

Submitted for the degree of Doctor of Philosophy

King's College London

May 2015

Abstract

Next generation sequencing generates a large quantity of sequence data which has the potential to be highly informative when evaluated using appropriate analytical methods. One of the key aims of human genetic disease studies is to use such methods to help identify sequence variants having some phenotypic effect. In the past few years, whole exome sequencing in particular has been used to identify single variants that cause many monogenic diseases. However, monogenic diseases in which genetic heterogeneity plays a role present a more difficult problem because different affected individuals in a study may not carry disease-causing mutations in the same gene.

A major focus of my work is to develop and implement algorithms to identify disease-causing variants in such diseases. In particular I make use of functional information, such as that encoded by interaction networks, to prioritise genes for follow-up analysis. In this thesis I present two different analysis tools designed for this purpose. Simulated datasets are constructed to demonstrate the utility of these tools and test their performance under varying conditions.

The tools are applied to a whole exome sequencing study for a genetically-heterogeneous monogenic disease (Adams-Oliver syndrome) with the aim of generating novel hypotheses regarding disease aetiology. This work also allows comparison and exploration of the challenges facing network-based methods in practice. The tools are also applied to a study of families exhibiting atypically strong recurrence of a complex disorder (Crohn's disease), testing the hypothesis that one or a small number of rare highly-penetrant variants might be implicated in each family. In this way it is proposed that the application of network-based methods to next generation sequencing data can help to describe disease mechanisms that move beyond monogenic diseases and towards more complex genetic architectures.

Copyright notice

Acknowledgements

I have had the great honour of working with four exceptional supervisors.

Firstly I would like to thank Thomas Schlitt for being prepared to support a supervisee with no previous bioinformatics or genetics experience, for helping me to establish my research and for countless enriching discussions that ranged from the daily practicalities to the elegant theoretical ideas of bioinformatics research.

I am no less indebted to Rebecca Oakey for her unhesitating commitment to see me through the second half of my PhD; I feel I am a much better scientist for Rebecca's experience and advice, and never once doubted that I was in safe hands.

I am also grateful to Reiner Schulz and Michael Simpson for agreeing to join my supervisory team. Reiner's insightful discussion and considered but ever-positive advice were of great benefit, while Michael provided a pragmatic approach and vital access to, as well as expert guidance in the analysis of, his exome sequencing data.

I would like to thank Laura Southgate for extensive support and discussion regarding the analysis of Adams-Oliver syndrome data. For the same support regarding the Crohn's disease work I would like to thank Natalie Prescott, along with Chris Mathew and Alex Onoufriadis.

I am grateful for the expertise and generosity of many other people in the Department of Medical and Molecular Genetics, including: Benjamin Lehne, Russel Sutherland and Inti Pedroso; Cathryn Lewis and Mike Weale; and Siobhan Hughes, Nikolaos Barkas and the other members of the Oakey/Schulz group(s). I also thank Andre Franke for hosting me at the University of Kiel, along with David Ellinghaus, Britt-Sabina Petersen and the other members of the Bioinformatics group.

As always I would like to thank my family for their unconditional support. Thank you also to Nick Bass and Li Chan for compelling me to consider a career in science and continuing to encourage and advise me. Most of all I would like to thank my wife, Ellie – for everything.

Abbreviations

ACC	Aplasia cutis congenita
AD	Autosomal dominant
AI	Acne inversa
AOS	Adams-Oliver syndrome
AR	Autosomal recessive
BIND	Biomolecular Interaction Network Database
BioGranat	Molecular Biology Graph Visualisation and Analysis Tool
bp	Base pairs
CCDS	Consensus coding (DNA) sequence
CD	Crohn's disease
COXPRESdb	Co-expression Database
CPDB	ConsensusPathDB
dbSNP	Database of Short Genetic Variation
EVS	[NHLBI Exome Sequencing Project] Exome Variant Server
FSS	Freeman-Sheldon syndrome
GBA	Guilt-by-association
GO	Gene Ontology
GWAS	Genome-wide association study
HGMD	Human Gene Mutation Database
HGNC	HUGO Gene Nomenclature Committee
HPRD	Human Protein Reference Database
HSP	Hereditary spastic paraplegias
IBD	Inflammatory bowel disease
ICMB	Institute of Clinical Molecular Biology [at the University of Kiel]
indel	Insertion or deletion
kbp	Kilo-base pairs
KCL	King's College London
LD	Linkage disequilibrium
Mb	Mega-bases
MCSC	Minimal connected set cover [problem]
MHC	Major histocompatibility complex

mRNA	Messenger RNA
MSC	Minimal set cover [problem]
MSigDB	Molecular Signatures Database
OMIM	Online Mendelian Inheritance in Man
PHA-I	Pseudohypoaldosteronism type I
PIN	Protein interaction network
PINA	Protein Interaction Network Analysis [database or network]
PPI	Protein-protein interaction
RGA	Region Growing Analysis
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPRING	SNV Prioritisation via the Integration of Genomic Data
TTLD	Terminal transverse limb defects
UC	Ulcerative colitis
WebGestalt	Web-based Gene Set Analysis Toolkit

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background to Genetic Disease

1.1.1 Genes and Disease

Many diseases have a genetic component. This means that there will be one or more positions in the genome at which variation in a person's DNA sequence can directly cause the disease, or can increase or decrease a person's likelihood of developing it relative to a person without that variation (all else being equal). In addition, genetic differences can affect the clinical phenotype and course of a disease (Zlotogora 2003), susceptibility to diseases caused by environmentally borne pathogens (Frodsham and Hill 2004) and response to treatment (Weinshilboum 2003). For most individual sequence variants that have a substantial effect on health, the harmful (disease-causing) allele will be less common in any given population than the non-disease allele; this is due to evolutionary pressure, where damaging sequence variants are selected against (Blekhman et al. 2008).

The central dogma of molecular biology describes how genetic information is transcribed from DNA into messenger RNA (mRNA), which is subsequently translated into a chain of amino acid residues to form the proteins that perform a vast range of cellular and inter-cellular functions (Crick 1970). There are several different types of genetic variation (den Dunnen and Antonarakis 2000). Of particular relevance to this thesis are DNA sequence variants affecting exons, the protein-coding DNA subunits of genes, which can alter downstream the ability of proteins to function correctly. If protein function is sufficiently impaired by the mutation this can cause or contribute to the clinical phenotypes characterising a disease.

Although there are examples of synonymous mutations playing a role in genetic disease (Sauna and Kimchi-Sarfaty 2011), most known disease-linked mutations in coding regions are non-synonymous, in that they alter the amino acid sequence and often consequently the structure of a protein (Cooper et al. 2010). *Single nucleotide variants (SNVs)* within exons are called *missense* if they result in an amino acid substitution or *nonsense* if they prematurely stop translation. *Splice-site mutations* at exon boundaries can cause exon-skipping or protein truncation. *Small insertions or deletions (indels)* that are in-frame (a multiple of three nucleotides long) affect the number of amino acids in a protein,

while frameshift indels result in a protein with a completely altered amino acid sequence which is also likely to be truncated or elongated.

Other types of genetic variation can cause disease, including sequence variants in introns or intergenic regions (Cooper et al. 2010) or larger-scale mutations including rearrangement or expansion of repeated sequences (Lupski 1998), large deletions or translocations (Abeysinghe et al. 2006) and chromosome number abnormalities (Torres et al. 2008).

However, recent advances in sequencing technology have made the identification of exonic sequence variants (particularly SNVs and small indels in non-repetitive DNA) in tens to hundreds of individuals fast and affordable, and therefore viable even for smaller research groups. Interpretation of this type of genetic variation is also more straightforward than non-coding variants. For these reasons, it is currently a key aim in medical genetics to identify exonic mutations that result in disease (Brunham and Hayden 2013). Knowledge of such variants can improve the understanding of the molecular basis of a disease, potentially leading to diagnostic and therapeutic advances. This provides strong motivation for the development of analysis methods for SNVs and small indels, which will be the focus of this thesis.

1.1.2 Monogenic Disease

In this thesis, the term *monogenic disease* will refer to diseases in which a single DNA sequence variant is sufficient to cause the disease. This avoids the ambiguity of the term *Mendelian disease*, since sequence variants involved in other types of genetic disease are also subject to Mendelian inheritance.

Monogenic diseases can result from sporadic (or *de novo*) mutations (errors in DNA replication after fertilisation) or from sequence variants inherited from one or both parents (Boycott et al. 2013).

Autosomal dominant (AD) monogenic diseases are caused by a single mutated copy of the disease gene (heterozygous mutation). When this is the result of an inherited sequence variant, one parent would be expected to have the disease phenotype. The mutations underlying AD disorders can be gain-of-function mutations in which the resulting protein gains some new abnormal function, or loss-of-function mutations if the gene is haploinsufficient (one copy alone produces insufficient protein product for normal function). Examples of AD disorders include Huntington's disease (gene *HTT*; Online Mendelian Inheritance in Man [OMIM] #143100) and Marfan syndrome (gene *FBN1*; OMIM #154700).

Autosomal recessive (AR) monogenic diseases require both copies of the disease gene to be mutated. This can result from homozygous mutations, where both copies contain the same sequence variant, or from compound heterozygous mutations, where different variants affect each copy. Frequently in AR disease cases, each parent will carry one mutated copy of the disease gene so that neither are affected by the disease, but one in four children will be (with an additional two in four also being unaffected carriers). Mutations underlying AR disorders are often loss-of-function for haplosufficient genes. Examples of AR disorders include cystic fibrosis (gene *CFTR*; OMIM #219700) and sickle-cell disease (gene *HBB*; OMIM #603903).

Mutations on the sex chromosomes can also result in monogenic disease. X-linked disorders can be dominant (such as Rett syndrome [OMIM #312750] caused by the gene *MECP2* and resulting in a severe phenotype in females but fatal in males) or recessive (such as Duchenne muscular dystrophy [OMIM #310200], caused by the gene *DMD* and much rarer in females than in males, who have no working second copy of the gene should they inherit a mutation).

There are estimated to be around 7,000 known monogenic diseases with the genetic basis of half of these having been determined (Boycott et al. 2013). There are likely to be many more monogenic diseases as-yet-uncharacterised due to being very rare or occurring in less well-studied populations (Antonarakis and Beckmann 2006).

1.1.3 Penetrance and Expressivity

For a monogenic disease, penetrance refers to the proportion of people carrying a disease-causing mutation that actually display the disease phenotype. There are examples of diseases for which penetrance is 100% (such as the AD disorder achondroplasia [OMIM #100800]), but often monogenic diseases can have incomplete penetrance.

A related concept is expressivity, which refers to the range and severity of phenotypic features that can occur in a monogenic disease patient (Lobo 2008). For example, the AD disorder neurofibromatosis, type I (OMIM #162200) can result in a range of clinical phenotypes, even within families where the same disease allele is causal.

There are several factors that can cause incomplete penetrance and variable expressivity. One is mutation type; for example in retinoblastoma (OMIM #180200) low penetrance disease-causing mutations in the gene *RB1* lead to partially functional or reduced levels of the retinoblastoma protein. However, other reasons can include environmental factors, epigenetic modifications and mutations in modifier genes (Zlotogora 2003; Cooper et al. 2013).

Modifier genes are perhaps the first step beyond the simple picture of monogenic disease presented so far. If the primary gene for a monogenic disease is the gene which can harbour a disease-causing mutation, mutations in modifier genes can also influence the disease phenotype by altering expression of the primary gene, or by playing some other functional role in the process impaired by a mutant primary gene product (Dipple and McCabe 2000; Weatherall 2001). Cystic fibrosis gives an example of how modifier genes can affect some of the clinical symptoms in affected individuals (Drumm et al. 2005).

1.1.4 Oligogenic Disease

Oligogenic disease refers to disorders in which mutations are required in two or more genes to produce a disease phenotype. There are several examples of diseases displaying digenic inheritance, requiring mutations in two genes (Schäffer 2013).

The difference between a digenic disease and a monogenic disease in which a modifier gene plays a role is blurry, since a modifier gene affects the presence or severity of the monogenic disease phenotype. While the distinction generally lies in the magnitude of effect of the primary gene (Samuels 2010), both cases illustrate that the genetic basis of disease is not necessarily simple. When multiple genes are involved in a disease it is more difficult to observe clear familial inheritance, and more difficult to identify disease-causing genes using traditional methods such as linkage and positional cloning.

Multiple genes can be involved in a disease because the protein encoded by each gene is only one element of a complex molecular system carrying out some physiological function. For example, one protein may directly interact with another protein, compensate for the lack of another protein, be involved in regulating expression of another gene, or play a role complementary to proteins in other tissues or systems. This complexity at the molecular level is one reason why it is rare to find straightforward genotype-phenotype relationships (Weatherall 2001). In fact diseases caused by one or a small number of genes lie at one end of a continuum from simple monogenic to complex polygenic disease (Antonarakis and Beckmann 2006).

1.1.5 Complex Disease

At the other end of the scale are common complex diseases such as diabetes or heart disease. A person's risk of developing a complex disease is linked to a large number of genetic and environmental factors, and they have no clear mode of inheritance (the elevated risk due to a family history of disease varies between complex disorders). Complex diseases tend to be more common than simple monogenic diseases because each underlying genetic

mutation has only a small effect on disease risk individually and is therefore not under the same selective pressure as a monogenic disease-causing mutation (Blekhman et al. 2008).

The common disease–common variant hypothesis proposed that the genetic mutations underlying complex diseases in an affected population are relatively common single nucleotide polymorphisms (SNPs) (Gibson 2011). Since 2005 (Klein et al. 2005), genome-wide association studies (GWAS) have used microarrays to genotype up to five million common SNPs (Illumina 2014b) in large case and control cohorts, testing for association of SNPs with disease state. The SNPs are spaced throughout the genome, making GWAS hypothesis-free. GWAS require large sample sizes to have the power to detect disease associations (Wellcome Trust Case Control Consortium 2007), and are subject to stringent multiple testing corrections (Johnson et al. 2010).

This approach has resulted in disease-associated loci being identified for many complex diseases (Hindorff et al. 2009). However, associated SNPs are unlikely to play a direct role in the disease themselves; more likely they are "tagging" variants with true disease involvement due to linkage disequilibrium (LD). Thus even replicated GWAS associations require much further work to draw conclusions about the molecular basis of a disease.

Despite their success as an analysis tool, GWAS have not fully explained the heritability of complex disease and it is therefore an ongoing problem to identify the role that rarer sequence variants play (Gibson 2011).

1.1.6 Genetic Heterogeneity

A key concept for this thesis is that of genetic heterogeneity, where several different sequence variants in the same gene (allelic heterogeneity), or sequence variants in several different genes (locus heterogeneity), can cause the same disease phenotype (McClellan and King 2010). Since allelic heterogeneity is compatible with a single disease-causing gene, this thesis will have a particular focus on locus heterogeneity, and it is primarily this type that will be meant when *genetic heterogeneity* is referred to.

It is important to distinguish between monogenic diseases with locus heterogeneity, which require a mutation in only one of several alternative genes (that is, there are several alternative primary disease genes), and oligogenic diseases, which require mutations in more than one gene. An example of locus heterogeneity is acne inversa (OMIM #142690), which can be caused by single sequence variants in any one of three genes involved in the γ-secretase complex (Wang et al. 2010a).

Motivated by recent reports attributing sporadic cases of relatively common neurodevelopmental disorders to *de novo* mutations in a wide range of genes, Gilissen *et al.*

suggest a relationship between the prevalence of a disorder and its mutational target size (Gilissen et al. 2011). This suggests that for more commonly observed genetic diseases causal variants might be found in many genes. Identifying and understanding these complex molecular mechanisms is a key goal of human genetic disease research, one for which next generation sequencing has provided renewed hope.

1.2 Next Generation Sequencing

1.2.1 The New Sequencing Technology

Initial draft sequences of the human genome were published in 2001, eleven years after the Human Genome Project commenced in 1990 (Lander et al. 2001; Venter et al. 2001). A finished sequence was declared complete in 2003 (National Human Genome Research Institute 2003). The availability of the complete (euchromatic) genome was a landmark development in genetics because it provided a comprehensive foundation for the sequencing of further genomes, for the study of functional elements such as genes and regulatory elements, and for comparative studies to investigate the genetic basis of disease.

However, the project cost an estimated $3 billion and took thirteen years to complete. Sequencing was completed using automated Sanger sequencing (Sanger et al. 1977), an enzyme-based sequencing method that could generate up to 115 kbp of sequence per day (Mardis 2011).

Since then, the study of genetics and genetic disease has been transformed by the arrival of next generation sequencing (NGS) methods. These are a group of commercialised high-throughput DNA (or RNA) sequencing methods that have dramatically improved the accessibility of genome-scale sequencing (Mardis 2011). Although the precise details vary between different NGS methods, a key innovation is the cyclic sequencing in parallel of millions of template DNA fragments, each of which is amplified to form a separate cluster on a solid surface. A single set of reagents is used to synthesise simultaneously the complementary DNA strands for each cluster, the sequence being determined using image-based detection when fluorescently-labelled nucleotides are incorporated (Shendure and Ji 2008; Metzker 2010).

Competition between providers of NGS methods has made them relatively affordable (Illumina are currently marketing a system capable of sequencing “the first $1000 genome” (Illumina 2014a)) and fast. Recent NGS instruments can sequence in the region of 10^{13} kbp per day (Mardis 2011). It is no surprise that such a radical improvement in

technology has changed the landscape of genetic disease research; studies are no longer limited to individual genes but can encompass the whole genome.

1.2.2 Read Alignment and Variant Calling

NGS results in relatively short DNA sequence reads of at most a few hundred base pairs (Metzker 2010; Liu et al. 2012). It is a considerable algorithmic challenge to map these reads to the reference genome of three billion base pairs, maintained by the Genome Reference Consortium. This problem is confounded by the fact that NGS reads can have an error rate of as much as 2% (Liu et al. 2012), and the fact that each individual's genome carries SNVs, indels and structural variation relative to the reference genome. Early alignment methods were based on hash-table algorithms (Flicek and Birney 2009); more recently methods such as Burrows-Wheeler Alignment (Li and Durbin 2009) make use of the Burrows-Wheeler transform for efficient read alignment (Flicek and Birney 2009; Li and Homer 2010).

Accurate sequencing requires each base pair to be read multiple times by different sequencing reads, and a genome-wide sequencing depth of 30× or higher is recommended (Koboldt et al. 2010; Ajay et al. 2011). Since reads need to be consistent, local re-alignment is performed to re-map reads where others indicate that an indel has occurred. For NGS methods which use paired-end reads, the placement of mate pair reads also needs to be consistent. Alignment algorithms are also able to take into account the sequencing quality scores at each base pair.

Read alignment is particularly difficult for repetitive regions (typically reads which map non-uniquely to the reference genome are discarded) or where there is structural variation relative to the reference genome. Performance varies depending on the tool used, parameters chosen and sequencing protocol used, but typically at least 10% (and sometimes substantially more) of the reads cannot be mapped to the reference genome (Hatem et al. 2013).

Variant calling is the process of determining where the sequenced genome differs from the (haploid) reference genome. A variant calling algorithm determines for each base pair whether a homozygous or heterozygous SNV has occurred, based on the likelihood of observing the mapped reads at that position. Mapping and sequencing quality are also taken into account (Li et al. 2008). Separate steps are required to identify indels and other structural variants. Since longer sequencing reads can be mapped with higher confidence they should result in better variant calling than shorter reads (Turner et al. 2009).

Since alignment and variant calling report the most likely genotype based on the sequencing data there will inevitably be false positive and false negative variant calls.

Several widely-used software packages provide universal tools for read alignment and variant calling, such as SAMtools (Li et al. 2009) and the Genome Analysis Toolkit (GATK) (McKenna et al. 2010). It is worth noting that several recent comparative analyses have shown that results can be highly dependent on the choice of software used (Liu et al. 2013; O'Rawe et al. 2013; Pabinger et al. 2014).

1.2.3 Whole Exome Sequencing

Rather than sequence the whole genome, a more cost-effective approach to genetic disease studies involves sequencing only the exome, the ~180,000 exons that comprise around 1% of the full genome sequence (Ng et al. 2009). Several methods exist for the targeted enrichment of genomic regions prior to sequencing. Most suitable for the whole exome is hybrid capture, in which a synthetic library of DNA fragments designed to cover the full exome sequence are used to hybridise to and capture complementary DNA from the input sample (Mamanova et al. 2010; Mertes et al. 2011).

For this purpose the definition of the exome is important. Most exome-capture kits work to a conservative definition based on the confirmed protein-coding sequences described in the Consensus Coding Sequence (CCDS) database (Pruitt et al. 2009), but a wider range (using definitions from databases such as RefSeq (Pruitt et al. 2014) or Ensembl (Flicek et al. 2014)) could include unverified coding sequences and potential pseudogenes (Ng et al. 2010c). For each exon, the targeted DNA sequence will usually include the exon body as well as short flanking regions at each end which incorporate splice acceptor and donor sites. Since some of the DNA fragments will overlap the target region boundaries it is also usual to co-capture longer flanking sequences of up to 200 bp (depending on sequencing read length) as a by-product; usually these are not utilised in genetic disease studies (Guo et al. 2012).

Whole exome sequencing nevertheless does not capture the whole exome. Exome-capture kits omit a small proportion of the exome from their target region due to technical difficulties in capturing those sequences. This can be up to 10% of CCDS-defined coding regions (Parla et al. 2011). In addition, due to sequence similarity with targeted regions a substantial fraction of sequenced reads will map to off-target regions; a recent comparison of exome sequencing platforms found a range of 9-35% (Clark et al. 2011). Coverage of the targeted regions depends on the depth at which sequencing is performed (i.e. the total number of reads), but the same study found that for 80M total reads only between 90% and 97% of targeted base pairs are covered at 10× depth. This is due in large part to GC-bias, where regions with high GC content systematically receive lower coverage (Clark et al. 2011).

Despite these limitations, the wide availability of NGS instruments and the relative affordability of whole exome sequencing has made it an extremely popular tool for genetic research, as will be discussed in detail in section 1.3 below.

1.2.4 Variant Databases

One important development that has occurred alongside the changes in sequencing technology has been the growth of databases making sequence and variant information quickly and easily available to researchers. There are a large number of valuable data resources (Fernández-Suárez et al. 2014), a few of which will be briefly mentioned here.

The 1000 Genomes Project aims to provide a comprehensive map of human sequence variation by using low-coverage sequencing to identify almost all variants with frequencies above 1% in a range of populations (1000 Genomes Project Consortium 2010). The UK10K Project has sequenced 4,000 genomes and 6,000 exomes from carefully phenotyped individuals, making the data available for the study of genotype-phenotype relationships (UK10K Consortium 2011). The Exome Variant Server (EVS) provides information on variants identified in more than 6,500 exomes (albeit with a particular focus on individuals with heart, lung and blood disorders) (NHLBI Exome Sequencing Project 2014).

The NCBI database of Short Genetic Variation (dbSNP; previously the Single Nucleotide Polymorphism Database) provides an archive to record all identified genetic variation across a range of species (Sherry et al. 2001). Newly identified variants can be submitted by users, and are associated with annotation including population frequency and disease relevance where available. More recently NCBI launched a companion database, ClinVar, which is more disease-focused and is designed to provide access to phenotype-related variant annotation (Landrum et al. 2014). Similarly, the Human Gene Mutation Database (HGMD) has a direct focus on disease, manually curating from published literature mutations that cause (or are associated with) genetic disease (Stenson et al. 2014).

Many journals now encourage authors to deposit newly identified sequence variants in a public database at the time of publication. This will ensure that comprehensive resources are available for use in methods development, epidemiological studies and in investigations into the genetic causes of specific diseases.

1.3 Genetic Disease in the Post-Genomic Era

1.3.1 Genetic Disease + NGS = Data

NGS is a powerful tool with which to study the genetic basis of disease because it can give a relatively complete picture of the genetic variation of an individual. However, with this clarity of vision comes a new problem: making sense of the deluge of information now available. For example, whole exome sequencing of an individual routinely identifies more than 20,000 SNVs alone (Bamshad et al. 2011). Thus in the past few years genetics has become more than ever a data science, with a great need for methods (including novel analysis strategies and new bioinformatics tools) that can pinpoint those variants playing a role in the disease under study.

1.3.2 Monogenic Disease: Intersection Filtering

A key milestone in this respect came in 2009, when Ng *et al.* at the University of Washington demonstrated a systematic approach to the identification of monogenic disease-causing sequence variants by whole exome sequencing (Ng et al. 2009). Their proof-of-principle study focused on Freeman-Sheldon Syndrome (FSS; OMIM #193700), a rare AD disorder affecting $\ll 1$ in 3,000 people and known to be caused by the gene *MYH3* (Toydemir et al. 2006). However, the analysis strategy is hypothesis-free, using no prior knowledge of the disease aetiology.

The authors sequenced the exomes of four unrelated individuals affected with FSS in order to look for genes in which all four carried a sequence variant. Unrelated individuals are used in order to minimise the number of identical-by-descent sequence variants, and therefore narrow the search space at the outset. The target capture region consisted of 26.6 Mb of CCDS coding sequence, which excluded 1.3 Mb of the exome that was considered poorly-mapped due to sequence similarity with other genomic regions. For each of the four individuals, between 95.9% and 96.4% of the target region was sequenced sufficiently to allow variant calling.

There were 2,479 genes in which all four affected individuals carried a non-synonymous coding SNV, splice-site SNV or coding indel (but not necessarily the same variant in all four individuals). In order to reduce the number of genes under consideration, several filtering steps were performed to discard variants that are less likely to cause a rare monogenic disease. The term *intersection filtering* has been used to describe this method (Robinson et al. 2011), which is illustrated in Figure 1.1.

Figure 1.1 – Overview of intersection filtering
An example of intersection filtering applied to four exomes using three filtering steps. (1) Exomes of four affected individuals, with sequence variants marked in red. Five genes (green boxes) contain variants in all four exomes. Variants listed in dbSNP (indicated in green) are filtered out (2). Excluded variants are denoted in grey; note that gene B no longer contains a variant in all four exomes after filtering. Subsequent filtering steps against control exomes (3) and using variant prediction tools (4) leave a single gene (E) with a variant in all four exomes. Gene E is a strong candidate for disease causality.

Firstly, all variants catalogued in dbSNP were excluded from the analysis. The reasoning for this is that a rare, highly-penetrant variant is expected to cause the disease and therefore a variant which causes FSS is unlikely to have been described in dbSNP without linking it to the FSS phenotype. Since 93-94% of the identified variants for each affected individual were in dbSNP this drastically reduced the number of genes in which all four individuals carried a variant to 53. Similar reasoning suggested that the disease-causing variant should not be present in any of eight healthy control exomes obtained from the HapMap project and sequenced by the authors using the same sequencing protocols as the FSS exomes. Further filtering of sequence variants on this basis left just a single gene with a sequence variant in all four affected individuals: the correct FSS gene, *MYH3*. This

demonstrates the power of intersection filtering to efficiently filter a large number of variants and quickly arrive at the disease-causing gene.

Had it been required, a third filtering step was also available. After excluding variants that were not predicted to be damaging by PolyPhen, a rule-based variant classifier which considers the amino acid change and its effect on protein structure (Ramensky et al. 2002), *MYH3* still contained a variant in all four affected individuals. Although not needed in this case the authors showed that this can still be an effective filtering step by considering the number of genes in which any three of the four FSS patients contained a sequence variant. After filtering against dbSNP and the eight HapMap individuals there were 22 such genes; this fell to three genes after further filtering using PolyPhen predictions. (Considering any three of four exomes allows the authors to conclude that "modelling of even a modest degree of genetic heterogeneity or data incompleteness is observed to have a significant impact on performance", which will be discussed further in section 1.3.11 below) (Ng et al. 2009).

1.3.3 Examples of Intersection Filtering

Following this pioneering initial paper, exome sequencing with intersection filtering quickly became a standard technique for the study of rare monogenic diseases. Since previous approaches to monogenic disease genetics relied on linkage studies within multiply-affected families, intersection filtering offered a new opportunity to study extremely rare disorders using only a small number of cases, which could include sporadic cases or familial cases where only one affected individual's DNA is available (Ku et al. 2011).

A number of review papers exhaustively list instances of monogenic disease gene identification through exome sequencing, the majority of which use some form of intersection filtering (Gilissen et al. 2011; Ku et al. 2011; Rabbani et al. 2012). Several examples will be discussed here to demonstrate variations of intersection filtering that have been successfully used in practice; these examples show that intersection filtering is an iterative investigative process in which logical filtering steps are applied that are consistent with prior expectations of the disease's genetic architecture based on its observed mode of inheritance, and with the clinical phenotypes observed.

Ng *et al.* followed their proof-of-concept FSS study with two applications to other diseases in 2010. Firstly they identified *DHODH* as the causal gene for Miller syndrome (OMIM #263750) by exome sequencing a pair of affected siblings and two unrelated cases (Ng et al. 2010b). Their study considered both AD and AR models, the recessive model requiring a gene contain two variants in each exome to be considered a candidate, and

ultimately identifying the causal gene. The filtering steps used were identical to the original FSS paper. The siblings were analysed no differently from unrelated individuals, but were advantageous because they were assured to have the same disease-causing gene, limiting any possible problems due to genetic heterogeneity.

Secondly the same group identified *MLL2* as a causal gene for Kabuki syndrome (OMIM #147920) by exome sequencing ten unrelated affected individuals (Ng et al. 2010a). Filtering was performed against known variants from dbSNP and the 1000 Genomes Project, and against 16 control exomes which included the FSS and Miller syndrome exomes. Interestingly the investigators ranked the exomes according to phenotypic severity and performed sequential intersection filtering with an additional filter for nonsense SNVs and frameshift indels only. Ultimately *MLL2* mutations were identified in nine of the ten exomes (seven were identified through whole exome sequencing and two via subsequent Sanger sequencing).

Fowler syndrome (OMIM #225790) is another example where mode of inheritance was successfully used as a filter (Lalonde et al. 2010). Using only two unrelated affected individuals, filtering against dbSNP and 1000 Genomes Project variants resulted in a single gene, *FLVCR2*, harbouring compound heterozygous mutations in both exomes.

Hoischen *et al.* found that the gene *SETBP1* causes Schinzel-Giedion syndrome (OMIM #269150) (Hoischen et al. 2010). By sequencing four unrelated affected exomes and filtering non-synonymous and splice site variants against dbSNP, they identified 12 genes harbouring variants in all four exomes. Ten of these genes were discarded because all four exomes carried the exact same variant, which were presumed to be as-yet-unidentified SNPs (that is, relatively common SNVs). Of the two remaining genes one was discarded because the investigators observed it had frequently contained variants in other in-house sequencing studies. The remaining gene, *SETBP1*, was found to be mutated in eight of nine additional cases. The disease-causing mutations were all located in a highly-conserved 11 bp exonic region and were confirmed to be *de novo* mutations by sequencing parental DNA.

MLL was identified as a causal gene for Wiedemann-Steiner syndrome (OMIM #605130) by intersection filtering on four individuals under an AD model, filtering against dbSNP, the 1000 Genomes Project and 600 control exomes (Jones et al. 2012). *MLL* was the only candidate gene remaining when looking for sharing between any three of the four exomes; no causal mutation was found in the fourth exome.

In a study of genitopatellar syndrome (OMIM #606170), the exomes of six unrelated affected individuals were sequenced, and intersection filtering identified causal variants in the gene *KAT6B* for five of them (Simpson et al. 2012). Initially an AR mode of inheritance was assumed, but no gene contained a post-filtering homozygous or compound heterozygous

variant in more than one exome. Causal variants were subsequently identified using a (sporadic) AD model.

Intersection filtering of five unrelated Floating-Harbor syndrome (OMIM #136140) exomes identified causal mutations in the gene *SRCAP* (Hood et al. 2012). Instead of filtering for novel variants only, this study used dbSNP and 1000 Genomes Project data to exclude variants with a minor-allele frequency of ≥ 1%, as well as filtering against 270 control exomes. The causal gene would have been identified using any four of the five exomes, and variants were subsequently found in *SRCAP* in all eight additional affected individuals tested.

In a study of mandibulofacial dysostosis with microcephaly (OMIM # 610536) undertaken by the same group, intersection filtering initially identified *MUC4* as the only gene containing post-filtering variants in all four sequenced exomes (Lines et al. 2012). However, this gene was discarded because it is frequently seen as a false positive result in exome sequencing studies of monogenic disease. Eight genes harboured variants in three of the four exomes, and the causal gene (*EFTUD2*) was discerned from these because it showed a region of reduced read depth in the fourth exome, which was subsequently proven to result from a chromosomal deletion spanning the last nine exons of the gene. Variants in *EFTUD2* were identified in eight out of eight follow-up cases; the range of mutation types found was consistent with haploinsufficiency causing this AD/sporadic disorder.

UVSSA was identified as a causal gene for the ultra-rare UV-sensitive syndrome (OMIM #614640) by intersection filtering the exomes of two unrelated affected Japanese individuals, neither of whom carried a mutation in *ERCC6* or *ERCC8*, the previously known causal genes (Nakazawa et al. 2012). The study assumed an AR disease model and *UVSSA* was the only candidate gene after filtering against dbSNP, 1000 Genomes Project and seven control exomes; variants were found in the same gene in a further Japanese case and an Israeli subject.

Polvi *et al.* identified *CTC1* as a causal gene for cerebroretinal microangiopathy with calcifications and cysts (OMIM #612199) by intersection filtering the exomes of four unrelated affected individuals (Polvi et al. 2012). Of the 15 subjects available to the investigators, they chose to sequence two pairs having the most similar clinical phenotypes in order to maximise the likelihood of the same gene being causal in each case. Under an AR model, filtering to exclude synonymous SNVs and those previously observed in dbSNP resulted in four genes having two or more variants in all four exomes. *CTC1* was highlighted as the most likely causal gene based on its function and the number of variants observed in a control set of exomes; variants were subsequently found in an additional eight unrelated patients.

Clearly then, intersection filtering is rarely as straightforward in practice as the FSS test case, and it must be applied carefully in order to make best use of the information obtained by exome sequencing and arrive at a sensible conclusion.

1.3.4 Key Assumptions of the Intersection Filtering Method

Although it is a powerful analysis method, intersection filtering makes several assumptions. Should any of these not hold, the disease-causing variant may not be identified. These assumptions include:

- ***That the disease-causing variants occur in exons***. If the disease under study is caused by an intronic or intergenic sequence variant it will not be captured by exome sequencing. A survey of HGMD in 2010 found that 11% of recorded disease-causing mutations occur in introns, at least some of these being >100 bp from the nearest exon and typically having a functional effect by activating a cryptic splice site resulting in aberrant splicing (Cooper et al. 2010). 3% occur in gene regulatory regions which may lie in untranslated regions at the gene boundary or considerably further upstream. The same article describes examples of diseases being caused by remote deletions and other genomic rearrangements, as well as by mutations in non-protein-coding genes. Nonetheless, exome sequencing remains a prudent approach because the majority of monogenic diseases-causing mutations identified to date *have* been found in exons.

  Successful identification of the disease-causing mutation can also depend on the definition of the exome used for DNA capture. For example in the Kabuki syndrome study (Ng et al. 2010a) the causal gene (*MLL2*) would not have been captured or sequenced using the CCDS definition of the exome, but fortunately the more inclusive RefSeq definition was used (Ng et al. 2010c).

- ***That the exonic variants are identified by whole exome sequencing***. As discussed in section 1.2.3 above, exome capture methods are unable to target the complete coding region, and the portion that is captured is subject to sequencing error. False negative variant calls, which could result in a true disease-causing mutation being missed, are a bigger problem than false positive calls, which can be identified by Sanger sequencing (Kuhlenbaumer et al. 2011). In the Kabuki syndrome study, two of the nine identified causal variants were missed by whole exome sequencing and only discovered subsequently by Sanger sequencing; fortunately the sample size was large enough that the disease-causing gene could still be identified by intersection filtering (Ng et al. 2010a).

- ***That the filters applied are appropriate***. At each stage of filtering the objective is to narrow down the list of remaining sequence variants in each exome to those most likely to cause a rare monogenic disease. Therefore there is a risk at each stage of filtering that the true disease-causing variant in one or more of the exomes might be incorrectly filtered out because it does not match closely enough the expected characteristics. For example, used blindly in the Miller syndrome study, PolyPhen variant effect prediction would have erroneously excluded a disease-causing variant in one of the exomes (Ng et al. 2010b). The next three points will discuss specific filtering assumptions.

- ***That the variant is non-synonymous***. Since synonymous mutations typically have a smaller effect than non-synonymous mutations it is often effective to filter them out, typically reducing the number of variants per exome by more than half (Bamshad et al. 2011). However, synonymous mutations can play a role in disease. While many known examples contribute to disease susceptibility, there are several diseases which can be directly caused by synonymous mutations, typically by introducing new splicing events (Sauna and Kimchi-Sarfaty 2011).

- ***That the mode of inheritance is as predicted***. If a disease appears to have AR mode of inheritance then a filtering step can be performed to leave homozygous or compound heterozygous variants only, as in the Fowler syndrome study (Lalonde et al. 2010). However, given that different variants can impact the encoded protein in different ways it is feasible that some sporadic cases could result from heterozygous changes.

- ***That the variant is rare and highly-penetrant***. This important assumption allows variants to be excluded by comparing against a population of healthy controls or database of known variants. If the disease under study has a severe clinical phenotype and is caused by a single mutation, it seems a reasonable assumption that the variant would have a severe molecular effect that (almost) always results in the disease. For example, the initial FSS study excluded variants which were observed in eight HapMap exomes, as well as those described in dbSNP (Ng et al. 2009).

  However, as described in section 1.1.3 above, it is possible that the true disease-causing variants are not fully penetrant for reasons that include modifier genes or environmental, epigenetic and other factors. A recent study investigated the presence in healthy exomes (both in-house and publicly available from the

1000 Genomes Project) of mutations listed in HGMD as causing early-onset dysmorphic disorders. Several examples were found, suggesting at least some of these variants show incomplete penetrance (Winand et al. 2014). These might therefore be observed in a set of healthy controls used for intersection filtering.

Arguably filtering against databases such as dbSNP will become less effective as more and more variants are identified and catalogued using NGS methods. The greater the number of whole exomes and whole genomes made publicly available, the more likely a rare mutation is to have been seen elsewhere and therefore a naive filtering approach that does not account for how those sequences were ascertained becomes less advisable.

An alternative to filtering based on entirely novel variants is to consider both novel and very rare variants, keeping for example variants to which the 1000 Genomes project or EVS ascribes an alternative allele frequency below a fixed threshold such as 1% or 0.1%. For example, a 1% threshold was used in the Floating-Harbor syndrome study (Hood et al. 2012). The limitations of requiring a very low population frequency are similar to requiring novel variants only.

- ***That a single gene is responsible for all or most cases of the disease***. Finally, a fundamental principle of intersection filtering is that by carrying DNA sequence variants with characteristics appropriate for involvement in a rare monogenic disease, unrelated individuals will independently incriminate a single functional unit (i.e. gene). If in fact locus heterogeneity is present, so that a number of alternative genes are responsible for the disease in the set of individuals sequenced, these may be indistinguishable from background variation because no gene causes a majority of cases.[*]

  Although intersection filtering studies are published which cannot identify a causal variant for a minority of the cases (e.g. Ng et al. 2010a; Simpson et al. 2012), it is difficult to estimate the number of studies which do not make it

[*] Technically intersection filtering can be effective *without* all or most cases of a disease being caused by a single gene: the method just requires that a gene is responsible for a sufficient number of cases to stand out against background variation after filtering. For small samples (as in the examples cited in section 1.3.3) this will usually mean one gene being responsible for all or most cases unless very stringent filtering can be employed. For example, Tatton-Brown *et al*. were able to show that mutations in *DNMT3A* cause an overgrowth syndrome with intellectual disability (OMIM #615879) because two of their ten cases carried a variant after filtering (Tatton-Brown et al. 2014). However, this was only possible because their "filtering" step was to identify the *de novo* sequence variants for each case, which requires the sequencing of additional exomes (discussed further in section 1.3.6). *DNMT3A* was also singled out due to a known relationship with another overgrowth predisposition gene.

to publication because they do not find *any* causal gene. However, it seems likely that the proportion of studies that are unsuccessful will grow as the number of diseases which are good candidates for intersection filtering (unsolved single-gene monogenic diseases; these are more likely to have a clear inheritance pattern and thus be studied first) falls.

Note that unlike association studies, which need to carefully match the ethnic backgrounds of case and control cohorts, intersection filtering does not require any explicit assumptions about the ethnic background of case and control exomes. Since genes have similar functions across ethnic populations, a mutation with a severe effect could generally be expected to have a similar phenotypic outcome for different ethnic populations. Control exomes in this case are used simply to rule out disease-causing variants. Nevertheless there are a couple of points to be aware of: different ethnic populations could in theory carry different alleles of modifier genes; and, any control exomes that have a different ethnic background to the cases may be less effective as filters because the set of SNVs found in the two populations will not fully overlap (although Ng *et al*. showed that the use of HapMap control exomes of different ethnic backgrounds could still be effective (Ng et al. 2009; Ng et al. 2010a; Ng et al. 2010b)).

It is important to be aware of these assumptions, and to factor them in to the study design and to the interpretation of intersection filtering results.

1.3.5 Some Genes Frequently Contain Variants

The reason that intersection filtering is needed at all is that every exome carries 20,000 or more SNVs (Bamshad et al. 2011), making it very difficult to pick out which one may cause a disease. One problem for the intersection filtering approach is that some genes are much more likely than others to contain rare non-synonymous SNVs, the majority of which are not disease-causing (Petrovski et al. 2013). These include genes such as *TTN* that have particularly long coding sequences, as well as highly polymorphic genes such as those encoding olfactory receptor proteins. Such genes cause problems because they tend to contain many SNVs per individual, giving a good chance that one or more will remain after all filtering steps are completed (e.g. Lines et al. 2012).

There are also technical reasons why genes might frequently appear as false positives in exome sequencing studies, including susceptibility to read misalignment and the presence of misleading information in the reference genome (Fuentes Fajardo et al. 2012).

Although there exist tools to predict whether variants are disease-causing (see section 1.3.7), coping with frequently-mutated genes continues to be a challenge for intersection filtering studies.

1.3.6 Monogenic Disease: Alternative Approaches

Whole exome intersection filtering is not the only approach to finding monogenic disease genes using NGS. Several alternative approaches take advantage of basic genetic principles to effectively pinpoint causal variants (often by building on pre-genomic techniques).

For X-linked disorders, it can be sufficient to sequence only the X chromosome and perform intersection filtering (Johnston et al. 2010). However, given the falling cost of NGS it may be more advisable to sequence the whole exome and exclude non-X chromosome variants as a provisional filtering step.

Where exomes are available from multiple affected family members it is possible to perform a combination of exome sequencing and linkage analysis. Since the same mutation is expected to cause the disorder for related individuals, only the variants shared by affected family members need be considered. The more recombination events that separate two members of a family (and the more family members sequenced), the shorter the list of shared variants will be; unaffected family members can also be sequenced to filter out non-causal mutations. Wang *et al.* provide an example where a disease-causing gene for a spinocerebellar ataxia (OMIM #613908) was identified by exome sequencing four affected individuals from the same family and filtering shared variants (Wang et al. 2010b). In other studies, linkage is performed first to identify a candidate region, with exome sequencing of related affected individuals used to pinpoint the causal gene within that region (Depienne et al. 2012; Rademakers et al. 2012). This approach also shows potential to study diseases with more complex genetics: recently, family-based exome sequencing was used to identify a risk-gene for familial Alzheimer's disease (Cruchaga et al. 2014).

For affected individuals born to a consanguineous union, homozygosity mapping can be used to identify long stretches of homozygous DNA sequence in a single patient. These autozygous regions are strong candidates to harbour mutations causing AR diseases. The mapping is typically done using genotyping arrays, before exome sequencing is used to examine the implicated regions in greater detail (e.g. Walsh et al. 2010; Abou Jamra et al. 2011; Cullinane et al. 2011; Erlich et al. 2011).

For disorders thought to be caused by *de novo* mutations (for example, where there is no family history of disease), exome sequencing of parent-child trios can be performed. The number of exonic mutations not carried by either parent is very small, typically 0-2

(O'Roak et al. 2011). This approach has been used successfully to identify causal mutations in multiple genes for relatively common sporadic diseases such as autism spectrum disorders (O'Roak et al. 2011; O'Roak et al. 2012) and schizophrenia (Fromer et al. 2014).

1.3.7 Variant Effect Prediction

A complementary approach to analysing large numbers of sequence variants is variant effect prediction. Tools that estimate the functional impact of a variant can be effective either as an intersection filtering step or as a method of prioritising genes for further study when multiple viable disease-causing variants remain after initial analysis. The aim of such tools is to estimate the likelihood that an observed variant is pathogenic. A range of evidence types can be used, such as the degree of evolutionary sequence conservation relative to other species or the impact of the variant on the encoded protein in terms of sequence (e.g. whether it occurs in a binding site), biochemistry (e.g. charge, hydrophobicity) or structure (e.g. where the variant occurs in the folded protein) (Cooper and Shendure 2011).

SIFT is a widely-used tool that scores amino acid-altering variants using sequence homology; variants in amino acids that are more highly conserved across a range of species are assumed to be more deleterious (Kumar et al. 2009). SIFT provides an example of a tool that directly estimates deleteriousness based on biological assumptions, as does MAPP, which incorporates biochemical properties (Stone and Sidow 2005).

Other tools use a machine-learning approach to estimate deleteriousness, basing predictions on a range of variant properties and using true positive and true negative datasets to derive classification rules. The advantage of this approach is that a wide range of variant properties can be considered without establishing in detail the biological relationships between them. Conversely, the interpretation of a variant's score is less clear.

For example, PolyPhen-2 predicts variant effect using eleven structure- and sequence-based features (including PSIC scores (Sunyaev et al. 1999), which use sequence data from homologous proteins to assess the likelihood of observing a variant) (Adzhubei et al. 2010). MutationTaster uses different classifiers for variants which cause no amino acid change (including intronic variants), variants affecting one amino acid and variants with more complex effect, and bases predictions on evolutionary conservation as well as biochemical and structural properties (Schwarz et al. 2010).

Several tools quantify evolutionary conservation of nucleotide sequence (rather than focusing on properties of the encoded protein), making them suitable for variants in both coding and non-coding regions. These include Genomic Evolutionary Rate Profiling (GERP), which identifies genomic regions containing fewer substitutions relative to other

mammalian genomes than expected (Cooper et al. 2005). PhyloP uses four different statistical tests (including GERP) to measure sequence conservation (and faster than neutral nucleotide substitution) (Pollard et al. 2010).

Given that these various approaches to predicting variant effect are based on different underlying assumptions, several attempts have been made to combine information from different approaches into an integrated pathogenicity score. Thus, for example, CAROL combines predictions from SIFT and PolyPhen-2 (Lopes et al. 2012), CoVEC integrates scores from four different individual classifiers (Frousios et al. 2013), and Condel from five (Gonzalez-Perez and Lopez-Bigas 2011). The KGGSeq analysis platform implements a logistic regression which combines scores from SIFT, PolyPhen-2, MutationTaster, PhyloP and a likelihood ratio test based on sequence conservation (Li et al. 2012). These tools show improved performance relative to any of the individual classifiers from which they are derived.

1.3.8 Verifying Causal Variants

When a monogenic disease-causing variant is identified by exome sequencing it is necessary to validate the finding and demonstrate supporting evidence for causality (MacArthur et al. 2014).

Given the relatively high error rate of NGS, the first requirement is to Sanger sequence the variant-containing gene in the case exomes, in order to confirm the mutations are true positives.

Variant effect prediction tools (discussed in the previous section) can provide supporting evidence that a mutation is expected to have a severe phenotypic effect. Beyond this, there are two broad classes of evidence to link the mutation to the *specific* phenotypic effect that is the disease.

Statistical genetic evidence can be provided by showing statistically significant association with the disease in a case-control study. Association could be tested for the individual variant by genotyping a large number of cases and controls at that locus, or for the containing gene by sequencing and performing a burden test. It is also possible to strongly suggest causality in familial cases by identifying segregating variants in the same gene in multiple affected families (Rabbani et al. 2012).

Experimental functional evidence can also apply at the variant or gene level, and can include: demonstrating that knocking out the gene or introducing the variant in an animal model results in a consistent phenotype; showing that the gene is expressed in disease-relevant tissues and/or that expression is affected by the variant; or demonstrating that the gene interacts with a gene known to result in the same or similar phenotype and/or that the

variant occurs in a protein domain likely to impact this interaction (for example, see Jones et al. 2012; Nakazawa et al. 2012).

1.3.9 Complex Disease: Rare Variant Association Testing

Beyond monogenic disease, exome sequencing is also used to study common complex disease, providing an alternative to the common disease–common variant study design of GWAS. NGS enables testing of the hypotheses that many small-effect rare variants or fewer large-effect rare variants explain the heritability of common diseases (Gibson 2011).

To this end, several region-based association tests have been developed that determine whether there is a statistically significant overoccurrence of rare variants within defined regions (usually genes) in case exomes relative to controls. Li and Leal compared single marker tests (which test single variants separately and then combine test results across a region), multiple marker tests (which test all variants in a region simultaneously using a multivariate test) and collapsing methods (which seek to avoid large multiple testing corrections by simply testing for presence or absence of rare variants in a region) (Li and Leal 2008). They developed a combined multivariate and collapsing (CMC) test and showed it to be more powerful than any of the individual methods. The CMC test is a "burden" test because it assumes all rare variants in a gene affect the phenotype in the same direction and to a similar degree (Lee et al. 2012).

Other tests have looked to improve power by using a weighting scheme that places higher emphasis on rarer variants (Madsen and Browning 2009), by using a variable frequency threshold to classify rare variants (Price et al. 2010) or by testing the variance rather than the mean of rare variant counts to allow for a mixture of risk and protective variants in a region (Neale et al. 2011). This last approach is generalised by the sequence kernel association test (SKAT), a widely-adopted method that allows covariates to be incorporated using multiple regression (Wu et al. 2011). Finally, since burden tests remain more powerful than SKAT when variant effects are highly correlated, SKAT-O seeks an optimal combination of the two tests (Lee et al. 2012).

These association tests can only test for the presence or absence of rare variants, and not whether those variants play a functional role. Thus their power is affected by the way that functional information is incorporated: collapsing tests suffer if non-functional variants are included (and potentially, due to sequencing gaps or annotation errors, if functional variants are excluded), while methods that can incorporate variant effect prediction (such as SKAT) rely on the accuracy of these scores. If a gene is significantly associated with a trait, association tests do not directly implicate individual variants and further work is required to

demonstrate the mechanism of causality – although it may be possible to pick out good candidate variants based on sequence conservation or functional effect (e.g. Cruchaga et al. 2014; Holmen et al. 2014).

1.3.10 Beyond Exome Sequence Data

In a sense, exome sequencing represents the tip of a data iceberg. Continuing improvements in NGS and other high-throughput technologies have led to a broad range of data-driven disciplines focusing on a range of physiological systems. One of the great challenges facing bioinformaticians in the immediate future is to keep up with the rate of data generation, developing novel integrative approaches to handling these data and drawing out biological insights (Hawkins et al. 2010).

It is expected that whole genome sequencing will ultimately become a more cost-effective approach to study disease than whole exome sequencing, allowing analysis of variants outside of protein-coding regions. However, identifying and understanding the disease-causing role of such variants will be difficult due to the size of the genome and our limited knowledge of non-exonic function; convenient functional units for intersection filtering are not well-defined as the genes are for the exome (Boycott et al. 2013). Whole genome sequencing of matched tumour and normal tissue pairs can also provide a complete picture of the mutations found in cancer cells (Goh et al. 2011), and an understanding of the regulatory function of non-coding regions is needed to help distinguish the key driver mutations that give cancer cells a selective advantage from the passenger mutations that occur as a by-product of uncontrolled cell division (Stratton et al. 2009).

However, annotation of non-coding regions is improving through efforts such as the Encyclopedia of DNA Elements (ENCODE) project, which has systematically searched the genome for units of DNA sequence with likely biochemical function, including transcribed regions, transcription factor binding sites, regions with distinct chromatin structures and enhancer- or promoter-like features, and DNA featuring specific histone modifications (ENCODE Project Consortium 2012). Although there are necessary limitations in the scope of this work (Eddy 2013), it is a good first step towards a more comprehensive understanding of the regulatory role played by non-exonic DNA.

Beyond sequencing an individual's DNA, NGS has enabled the acquisition of a number of high-throughput data types. This is frequently referred to as "omics" data when it encompasses all of a particular class of molecule. For example, RNA-seq has become a viable alternative to expression microarrays for the study of gene expression, allowing quantification of the whole *transcriptome* in a given cell type under different conditions (Wang et al. 2009).

Likewise, data generation has begun apace in fields such as: *epigenomics*, which studies DNA modifications (for example, bisulfite sequencing can identify genome-wide DNA methylation (Krueger et al. 2012)); *proteomics*, the study of the structure and function of the entire set of proteins in an organism (Altelaar et al. 2013); *metabolomics*, in which chemical processes in cells are studied via metabolites (Wishart et al. 2013), and *metagenomics*, which studies environmental DNA (for example characterising the essential microbes found in the human gut by their genomes (Qin et al. 2010)). This thesis will have a particular interest in the *interactome*, as will be discussed in detail in section 1.4.

1.3.11 The Need for New Bioinformatics Methods

While the growing volume of biological data provides new opportunities to improve our understanding of human genetic disease mechanisms, it will be a challenging process to integrate such diverse data types and draw meaningful conclusions.

For monogenic disease in particular, identifying the causal variants from whole exome sequencing data will become increasingly challenging as intersection filtering and related approaches provide diminishing returns. A variety of complicating factors mean that the continued success of existing methods will be limited. These include locus heterogeneity, suggesting a more complex underlying biology and increasingly likely as more common genetic diseases are studied (Gilissen et al. 2011), the existence of diseases with overlapping clinical phenotypes, and phenocopies which may confound analyses because they have different disease mechanisms.

New methods that can integrate emerging omics data into whole exome sequencing studies should prove more successful at untangling these complex genotype-phenotype relationships. The goal of this thesis will be to develop and examine exome sequence analysis methods that counter genetic heterogeneity by using interaction networks to identify putative molecular pathways underlying monogenic disease.

1.4 Interaction Networks

1.4.1 Network Basics

Networks provide a structured way to represent information about how genes interact with one another. Genes are represented as network nodes, and two nodes are connected by an edge when the two genes are related, the type of relationship defining the type of network. Edges can be directed and/or weighted, depending on the information they express (see Figure 1.2).

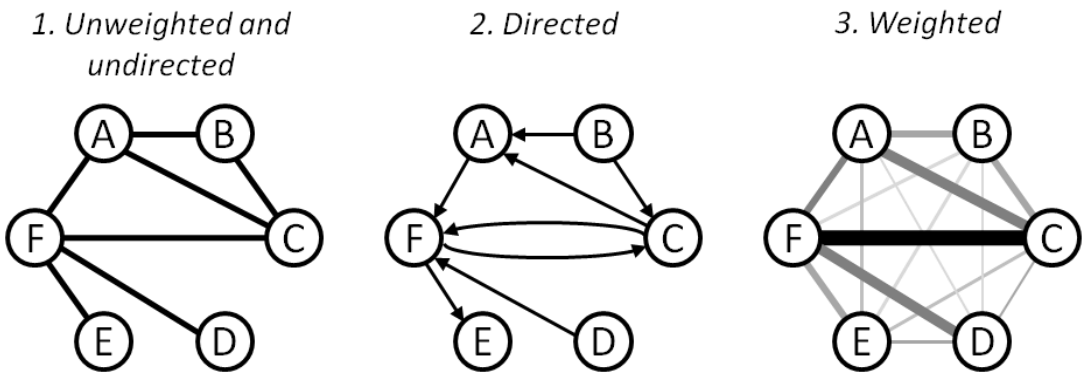


Figure 1.2 – Network edge types
1. Unweighted and undirected edges represent binary relationships between genes such as interaction of their protein products (A interacts with B but not with E); 2. Directed edges can indicate a directional relationship such as regulation; 3. Weighted edges can be used to indicate correlation or interaction confidence (here edges get thicker and darker with higher weight).

In mathematical terms interaction networks are *graphs*. Graph theory has a long history as a well-studied branch of discrete mathematics; this means that many network properties are well characterised and that efficient algorithms exist for many standard network analysis problems. From a biological point of view, graph theory provides a powerful framework which can take pairwise relationships between genes and suggest hypotheses about the way those genes function together as part of a system (Mason and Verwoerd 2007). Networks can be employed to systematically interpret experimental data for a wide range of investigations, including the study of disease processes.

Graph theory has a wealth of terminology and it is worth defining some basic terms here (illustrated in Figure 1.3). A *neighbour* of node *g* is any node that is connected to *g* by an edge, and the number of neighbours is denoted the *degree* of *g*. A path from node *g* to node *h* in an unweighted and undirected network is a sequence of nodes $g = g_1, g_2, \ldots, g_n = h$ such that successive nodes in the sequence are neighbours. A path's length is the number of edges it traverses, and the *minimum distance* between node *g* and node *h* is the length of the shortest path that connects them. Two nodes are in different *components* of a network if there is no path that connects them. *Cliques* are subnetworks in which every node is connected to every other node by an edge.

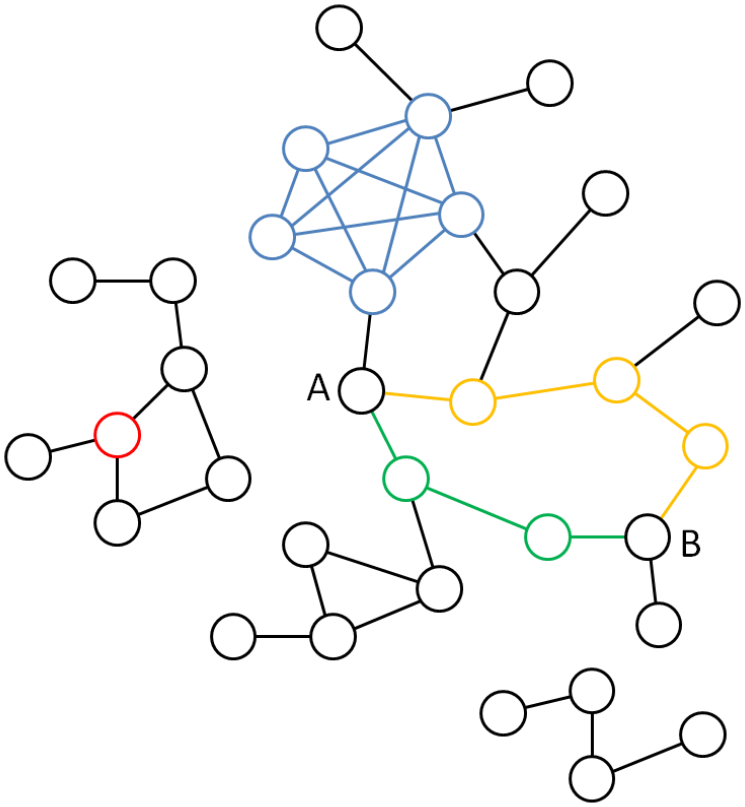


Figure 1.3 – Network terminology
This network consists of three components; no path exists from a node in one component to a node in another. The red node has three neighbours and hence degree 3. A clique of five nodes is indicated in blue. The yellow nodes and edges indicate a path between nodes A and B of length 4. However the minimum distance from node A to node B is 3, as indicated by the green path.

1.4.2 Protein Interaction Networks

Protein interaction networks (PINs) connect two genes when there is evidence that the proteins they encode interact physically* (Raman 2010). Two high-throughput methods have mainly been used to generate evidence of protein-protein interactions (PPIs) (Lehne and Schlitt 2009).

The yeast two-hybrid method involves binding protein A to the transactivation domain and protein B to the DNA-binding domain of the yeast transcription factor GAL4. If the transactivation domain is located in a promoter region of a gene in yeast it will trigger its transcription. Interaction of proteins A and B thus allows the transactivation domain to be

* Formally a PIN connects proteins that interact, not genes. However, interactions between specific protein isoforms are not readily available in protein-protein interaction databases and in general PINs treat all isoforms of a protein as a single node (in which case it is convenient to refer to this node by the encoding gene). A recent report of a small isoform-resolved PIN suggests that it will become necessary to avoid this conflation in future (Corominas et al. 2014).

recruited to an upstream activating sequence inserted in front of a reporter gene, resulting in its expression (Fields and Song 1989).

The tandem affinity purification and mass spectrometry (TAP-MS) method involves tagging the protein of interest with a compound protein fragment that allows two rounds of affinity purification. Proteins interacting with the tagged protein are co-purified and subsequently retained when the tag is cleaved. After eluting, the proteins making up the complex are identified via mass spectrometry (Rigaut et al. 1999).

Several databases collate interactions derived via these high-throughput methods, along with literature-curated interactions identified in smaller-scale experiments and interactions deposited directly by researchers. PPIs are studied for a wide range of organisms; larger databases that contain human PPIs include the Human Protein Reference Database (HPRD; Keshava Prasad et al. 2009), BioGRID (Chatr-Aryamontri et al. 2013), IntAct (Orchard et al. 2014) and the Biomolecular Interaction Network Database (BIND; Isserlin et al. 2011).

There also exist meta-databases, such as the Protein Interaction Network Analysis (PINA) platform (Cowley et al. 2012), which integrate data from several of these primary sources. Although meta-databases can include more data they often lag behind the primary databases in incorporating new interactions. It is also a non-trivial task to integrate different data sources, for example due to the complex mapping between different gene and protein identifiers (Lehne and Schlitt 2009).

PINs can be derived by downloading the contents of a PPI database and constructing an unweighted and undirected network based on pairwise PPIs.

1.4.3 Other Types of Interaction Network

Although PINs are perhaps the most studied type of biological network in the context of human disease, networks can be constructed using any type of relationship between genes. Many other classes of network have been studied.

Co-expression networks connect two genes when there is statistically significant similarity in their expression patterns across gene expression datasets covering multiple cell types and conditions (Stuart et al. 2003). Genome-wide expression profiles can be obtained using microarrays comprising DNA probes; the mRNA present in a cell is converted to complementary DNA (cDNA), labelled with a fluorophore and hybridised to the array, so that an optical signal will be emitted at probes corresponding to expressed mRNA (Schena et al. 1995). More recently, RNA-seq has provided a sequencing-based approach to mRNA quantification (Marioni et al. 2008). One useful resource is COXPRESdb (Co-expression Database), a database that provides co-expression relationships derived from microarray

experiments for over 19,000 human genes (Obayashi et al. 2013). Co-expression networks are undirected but edge weights can express the degree to which each pair of genes is co-expressed. An unweighted (binary) network can be derived using a threshold for correlation.

Gene regulatory networks have directed edges that describe regulatory relationships between genes. Transcription factor proteins regulate the expression of other genes by binding to their promoter regions, either enabling or preventing the recruitment of RNA polymerase to begin transcription. Hence a network encoding such relationships can contain different node types (transcription factor genes and regulated genes) and directed edges (Lee et al. 2002; Shen-Orr et al. 2002; Babu et al. 2004).

Metabolic networks connect enzymes that catalyse consecutive reactions and correspond to implicit underlying networks of chemical reactions with nodes representing biological compounds and edges representing reactions (Duarte et al. 2007; Yamada and Bork 2009).

Networks of *genetic interactions* can also be inferred by knocking out genes in model organisms, for example by comparing genome-wide expression levels (Li et al. 2013) or phenotypic traits between different gene knockouts (Wang et al. 2013b), or looking for deviations from the expected effect of double-knockout models based on the observed effect of single gene knockouts (Costanzo et al. 2010).

Literature or *co-citation networks* are undirected networks that are constructed computationally and do not directly use experimental results. In the simplest case, two genes are connected when they are found to have been cited together in publication abstracts significantly frequently, suggestive of some functional relationship (Jenssen et al. 2001). More sophisticated text-mining approaches exist to identify connections between genes, such as the "relatedness" measure developed by Raychaudhuri *et al.* that assesses the similarity of descriptive text in literature abstracts without necessarily requiring co-citation (Raychaudhuri et al. 2009).

However, due to the lack of direct empirical evidence supporting specific interactions literature networks are not usually depended on exclusively for the study of molecular systems or the interpretation of experimental data; rather, text-mined evidence for interaction is used to improve the accuracy of other types of network, for example by providing prior knowledge from which Bayesian networks describing co-expression (Djebbari and Quackenbush 2008) or gene interaction (Olsen et al. 2014) can be constructed. STRING is a database of PPIs which supplements interactions derived from high-throughput experiments with interactions predicted by text-mining, as well as those predicted using evolutionary arguments by comparison with other organisms (Franceschini et al. 2013).

Finally, with the goal of achieving a more comprehensive picture of sub-cellular-level processes, functional networks have been constructed which integrate diverse interaction types. For example, the HumanNet network connects genes using functional relationships learned from PPIs, co-expression, genomic context of orthologous genes across archaea and bacteria, protein domain co-occurrence, interaction in model organisms and literature-mining (Lee et al. 2011). The Multinet network combines interactions from protein interaction, genetic, regulatory and metabolic networks together with phosphorylation relationships and membership of common signalling pathways (Khurana et al. 2013).

Interaction networks are typically used to provide a tractable genome-wide representation of our knowledge of how genes relate to one another, giving a context for the further analysis of many types of experimental results (discussed in subsequent sections). When deciding on an interaction network for this purpose it should be noted that there is a natural conflict between genomic coverage on the one side, and interpretability of, and confidence in, the interactions represented on the other.

1.4.4 Network Topology Provides Clues to Gene Function

It is well established that interaction networks are organised non-randomly, so the study of a network's global structure or topology can reveal clues to the underlying biology of the genes involved.

Interaction networks tend to display a scale-free topology, which means the proportion of network nodes having degree k tends to have a power-law distribution, such that $P(k) \approx Ak^{-\gamma}$ for some degree exponent γ (usually between 2 and 3) and constant of normalisation A (Albert 2005). In practice this means that interaction networks contain many nodes of small degree and fewer nodes of large degree, with a handful of very high-degree "hub" nodes (see Figure 1.4). (Hub genes can be alternatively defined as those having a high *betweenness centrality*, where a large proportion of the shortest paths between all pairs of nodes pass through the gene, or *closeness centrality*, where the gene has a low average distance to all other nodes.)

Studies in model organisms have shown that hubs in PINs tend to represent essential genes for which knockouts are lethal (Jeong et al. 2001; Hahn and Kern 2005), and to be evolutionarily conserved (Fraser et al. 2002; Saeed and Deane 2006; Kim et al. 2007). This is generally attributed to the central role played by hub genes in the network's structure (removal of hub genes has more severe consequences to network topology than non-hub genes), but an alternative theory is that it is not genes but certain PPIs that are essential, with hub genes the most likely to participate in an essential PPI because they participate in the most PPIs (He and Zhang 2006). Hub genes can be further classified and interpreted as

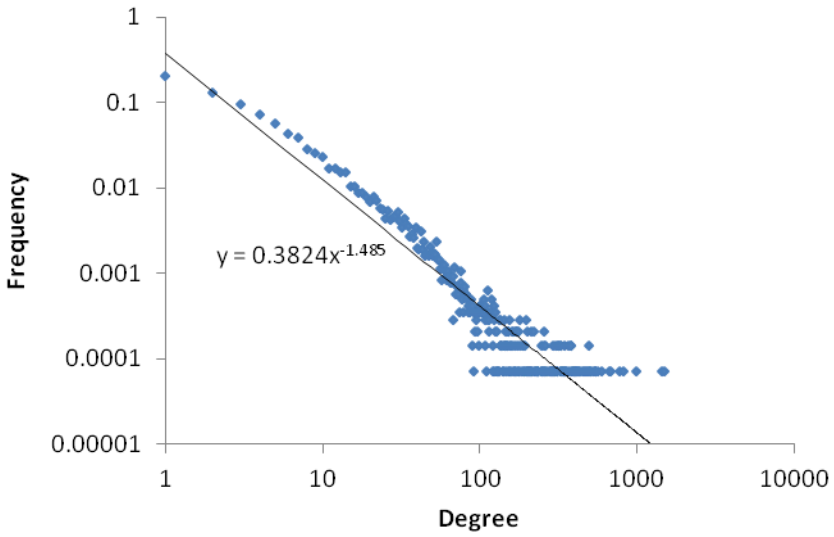


Figure 1.4 – Example degree distribution for a scale-free network
The Multinet network that combines various interaction types (see section 1.4.3) has an approximately scale-free topology. Degrees plotted in blue, trendline shown in black.

"party" or "date" hubs depending on whether they interact with many other genes simultaneously (playing a central role in a specific function) or at different times and cellular locations (connecting genes involved in different processes) (Han et al. 2004). The relative essentiality of party and date hubs has been debated (Han et al. 2004; Kim et al. 2006).

Another area of study is how networks are organised into distinct functional modules. Communities or modules can be loosely defined as groups of network nodes that are connected by relatively many edges but have relatively few connections to nodes outside of the group (Fortunato 2010) (see Figure 1.5). Many clustering algorithms exist to determine communities from global network structure, and applications of such algorithms to interaction networks have revealed communities of interacting genes that are shown to have related functions, for example by comparison with known protein complexes or Gene Ontology terms (Spirin and Mirny 2003; Luo et al. 2007; Lewis et al. 2010; Li et al. 2010).

Weighted gene co-expression network analysis has emerged as a popular method of analysing gene expression data by systematically identifying modules of genes that have correlated expression (Langfelder and Horvath 2008). Modules are then considered as basic functional units to simplify the analysis of external data, and to generate systems-level hypotheses by studying inter-module relationships.

It has also been demonstrated that network structure on a local, rather than global, scale can be informative. Network motifs, small recurring patterns of interaction that occur more frequently than expected by chance, have been identified in many types of directed networks (Milo et al. 2002). In gene regulatory networks, motifs correspond to specific types

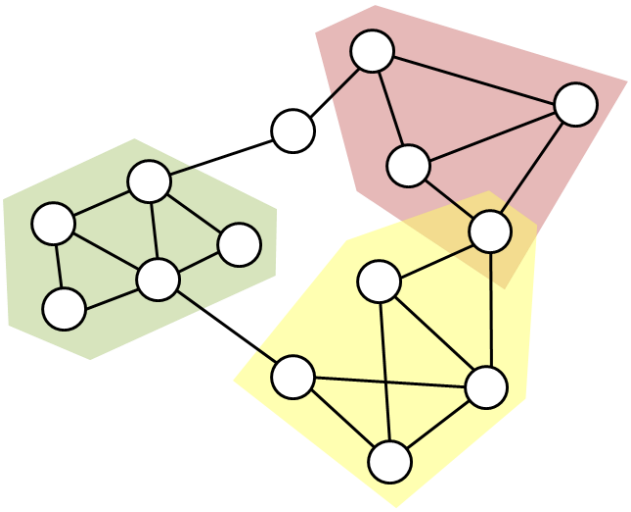


Figure 1.5 – Communities in a network
Communities consist of nodes that are more highly connected to each other than to nodes elsewhere in the network; communities can overlap.

of regulatory mechanism (Babu et al. 2004) and almost all regulation can be understood in terms of a small number of motifs (Alon 2007). Many algorithms have been developed for motif identification but it has been noted that caution should be used in their application because not all motifs in a network are necessarily biologically meaningful (Kim et al. 2011).

To examine the local- to intermediate-range structure of interaction networks, Pržulj *et al.* developed the concept of graphlets, which characterise all possible induced subnetworks of up to five nodes (Pržulj et al. 2004). For each gene a "graphlet degree signature" is determined, describing the number of graphlets of each type in which it participates. Graphlet degree signatures have been shown to correlate with gene function (Milenković and Pržulj 2008) and to highlight biologically important genes (Milenković et al. 2011).

But while it has been shown that network topology can be informative regarding gene function, an overreliance on such analyses has been cautioned. For example, Gillis and Pavlidis suggest that the presence of multifunctional genes may give a deceptively favourable impression of consistent function in local network regions (Gillis and Pavlidis 2011). Meanwhile Ideker and Krogan highlight that inferences drawn from static networks (as described in sections 1.4.2 and 1.4.3) are limited, and that consideration of how interaction networks change in response to different biological stimuli is necessary for a fuller understanding of how genes function together (Ideker and Krogan 2012).

1.4.5 Network Properties of Disease Genes

By considering genes that contain known inherited disease mutations, Feldman *et al.* show that disease-causing genes tend to be of intermediate connectivity in PINs, having higher average degree than non-disease genes but generally not representing hubs (Feldman et al. 2008). Goh *et al.* distinguish essential from non-essential disease genes and find no tendency for the non-essential disease genes to be network hubs (Goh et al. 2007). These findings are consistent with the notion that mutations in hub genes are more likely to be lethal during gestation, while mutations in genes of intermediate centrality are more likely to be viable, but result in a disease phenotype (Goh et al. 2007; Feldman et al. 2008). In the Multinet network, which integrates various types of interaction, genes were categorised into tolerant of loss-of-function mutations, neutral, disease-causing and essential, and a statistically significant increasing trend was observed in average degree across these categories (Khurana et al. 2013).

It has been shown that disease genes are significantly more likely to interact with one another than with non-disease genes (Gandhi et al. 2006). Further, when OMIM diseases were classified based on organ system and disease type, most classes exhibited strong enrichment for PPIs between within-class causal genes relative to across-class causal genes (Gandhi et al. 2006). Likewise, pairs of genes causing the same disorder were found to be significantly likely to interact in a PIN (Goh et al. 2007; Feldman et al. 2008). Van Driel *et al.* use a text-mining framework to quantify the phenotypic similarity between different diseases, finding that this measure is correlated with the connectedness of causal genes in a PIN (van Driel et al. 2006). This supports the idea of moving past arbitrary syndromic definitions, considering instead that similar clinical phenotypes might result from mutations in a localised region of an interaction network (Oti and Brunner 2007).

It could be argued that the reported enrichment of interactions between disease-causing genes results from a knowledge bias, since the functional roles of such genes are closely scrutinised once they are identified. However, this argument is countered by the finding that genes associated with the same complex disease via GWAS, a hypothesis-free approach, also display an enrichment of interactions (Barrenas et al. 2009). Similar evidence has been identified by exome sequencing. Two studies found that a statistically significant number of *de novo* truncating or severe missense mutations identified in autism spectrum disorder trios occurred in genes forming a highly interconnected subnetwork in a PIN (Neale et al. 2012; O'Roak et al. 2012). Recently, candidate causal genes for hereditary spastic paraplegias (HSP) found by intersection filtering were also found to be significantly highly connected (Novarino et al. 2014).

Barabasi *et al.* provide a neat conceptualisation of this relationship. They differentiate topological modules from functional modules (connected sets of genes with related function, discussed above in section 1.4.4) and disease modules (connected sets of genes with related disease roles), suggesting that a degree of overlap between these module types can be usefully assumed provided the distinction is not forgotten (Barabasi et al. 2011). A more detailed understanding of the topology of such functional and disease modules is needed to explain how diseases can display phenotypic variability, and how genes can have roles in different diseases; for example genetic variants which affect different binding domains of a protein can "perturb" different edges of a PIN, with distinct phenotypic consequences, each different again to a variant which causes a complete loss of function, effectively removing a node from the network (Zhong et al. 2009).

1.4.6 Network Methods for Disease Gene and Pathway Identification

Given that networks describe functional relationships between genes and that disease-causing genes tend to be co-localised in interaction networks, numerous methods have been developed that make use of network data to identify or prioritise disease-causing genes in high-throughput studies.

To investigate how well network-based guilt-by-association (GBA) methods can predict disease-causing genes, Lee *et al.* considered six different GBA methods applied to the HumanNet confidence-weighted network of multiple interaction types, with performance measured using cross-validation analysis on sets of known disease genes (Lee et al. 2011). Starting with initial gene scores of 1 for disease genes and 0 for all other network genes, two diffusion methods that are mathematically related to Google's PageRank algorithm (Brin and Page 1998) offered the best performance. Other methods considered were based on simple neighbour counting, naive Bayes label propagation (in which edge weights, rather than the neighbouring genes themselves, are summed), network clustering and an electrical circuit algorithm (Lee et al. 2011).

However, most *de novo pathway discovery* methods are designed for application to complex diseases and cancers (Lehne and Schlitt 2012). Several approaches focus on expression microarray data, with the aim of identifying *dysregulated subnetworks* – connected sets of genes in the network that collectively show differential expression between two experimental conditions (for example different samples representing different disease states). Ideker *et al.* developed jActiveModules, a tool which uses simulated annealing (Kirkpatrick et al. 1983) to identify subnetworks that minimise a combined differential expression p-value function (Ideker et al. 2002). The MATISSE tool supplements the edges of a PIN with artificial connections denoting similarity in the

expression dataset being analysed, before clustering to identify relevant network modules (Ulitsky and Shamir 2007).

A related tool from the same group, DEGAS, reformulates the problem, considering separately the genes differentially expressed in each of a number of disease-affected individuals (Ulitsky et al. 2010). Within a PIN they search for the smallest connected subnetwork in which all but *l* of the individuals have at least *k* differentially expressed genes, for suitable parameters *k* and *l*, employing a greedy algorithm which is limited to subnetworks of small radius around a root node. The KeyPathwayMiner tool addresses the same problem, taking a less sophisticated approach by classifying genes as differentially expressed if they exhibit differential expression in all but *l* samples, and seeking maximal subnetworks in which all but *k* genes are differentially expressed (Alcaraz et al. 2011; Alcaraz et al. 2012). Subnetworks are identified using an Ant Colony Optimisation heuristic, and results are broadly comparable to DEGAS (but the authors note that the repurposed *k* and *l* parameters are more intuitively interpretable and it is no longer necessary to specify a minimum subnetwork size).

Several approaches seek to identify subnetwork markers whose dysregulation is predictive of cancer type or prognosis; various methods have been proposed based on graph-search (Chuang et al. 2007; Dao et al. 2010; Chowdhury et al. 2011) and integer linear programming algorithms (Dittrich et al. 2008; Backes et al. 2012).

Besides expression microarray data, the other main application of networks in disease gene identification has been to GWAS data. Networks have been employed both to highlight potentially relevant genes from the many which can be in LD with genome-wide significant SNPs, and to explore the disease role of SNPs that show nominally significant disease association but do not quite exceed the stringent genome-wide significance level after correction for multiple testing.

In the former category, Rossin *et al.*'s DAPPLE tool seeks to identify disease genes and suggest underlying molecular processes by identifying connected genes in a PIN that are in LD with different disease-associated GWAS SNPs. Connectivity of identified subnetworks is assessed by comparison to 50,000 randomly permuted PINs (Rossin et al. 2011). This approach is conceptually similar to GRAIL (developed by the same group), which examines literature-derived relationships instead of PPIs to infer disease-relevant interactions between genes in different implicated GWAS loci (Raychaudhuri et al. 2009).

On the other hand, to overcome the stringent genome-wide significance cut-off required for GWAS, Baranzini *et al.* present an analysis strategy which searches for subnetworks of a PIN showing disease association. Gene-wise p-values are obtained by taking the most significantly associated GWAS SNP within each gene, and the

jActiveModules tool (described above) is employed using these gene-wise measures of association in place of the usual differential expression p-values (Baranzini et al. 2009). Region Growing Analysis (RGA) is a related network-search method. It first ranks genes according to GWAS SNP association (adjusted for gene size), before examining various ranking thresholds to identify network regions enriched for GWAS signal. Significance is established using degree-constrained network permutation (Lehne 2011). In the HumanNet edge-weighted network, Bayes label propagation (as discussed above) was shown to successfully identify disease-linked genes which interact with those identified via GWAS (Lee et al. 2011).

To date, there have been relatively few network methods designed specifically to identify disease-causing genes using NGS data. Whole exome sequence data usually contains few directly disease-relevant variants and much noise, and initially, at least, sample sizes have been relatively small. Therefore it may not be possible or appropriate to directly apply network methods designed for microarray or GWAS data, which have different statistical properties. The secondary nature of network-based methods means that their development will inevitably lag behind direct analysis of NGS data, particularly when it has been initially so fruitful.

However, for diseases with known causal genes, one natural approach is to use simple GBA to help prioritise variants identified by whole exome sequencing. Thus tools have begun to appear which provide an integrated exome analysis framework which can perform variant filtering and highlight those variants found in genes which interact with some specified seed list (Li et al. 2012; Sifrim et al. 2012). One successful application of GBA to exome data was the HSP study referred to in section 1.4.5 above; after initial exome sequence analysis identified a subnetwork of causal genes, examination of the neighbouring genes and a second cohort of exomes identified three further candidates (Novarino et al. 2014).

An example where an existing tool could be applied is one of the autism spectrum disorder studies, where the DAPPLE tool (originally designed to analyse GWAS loci) was used to infer a relevant PIN subnetwork from *de novo* mutations obtained by sequencing trios (Neale et al. 2012). The enrichment for interactions between genes harbouring mutations in different families lends credibility to the assertion that these genes underlie autism.

A very recently published tool, SPRING (SNV Prioritisation via the Integration of Genomic Data) aims to prioritise non-synonymous SNVs found in a single exome for disease causality, by integrating several data sources including functional prediction scores to assess likely pathogenicity and disease-relevance measures for the containing genes. One

disease-relevance measure is proximity in a PIN to a set of seed genes; where known disease-causing genes do not exist these are generated from genes that cause phenotypically similar disorders (Wu et al. 2014).

Outside of the monogenic disease field, VarWalker is designed to identify consensus mutation networks in cancer, with the aim of identifying driver mutations in NGS data. In each sample, mutation-harbouring genes are identified from sequence data, before random walk with restart in a PIN identifies interacting genes (with those that are frequently found in permuted networks being removed). Consensus mutation networks are found by comparing across samples (Jia and Zhao 2014). For complex diseases, the first applications of networks to rare variant association tests are being reported. Wu and Zhi compare pathway-based approaches for sequence-based association tests (Wu and Zhi 2013), which could include putative pathways identified in interaction networks; this approach was taken in a recent study of type 2 diabetes, where neighbours of known diabetes genes were used to define pathways (with negative results) (Lohmueller et al. 2013).

While network methods have long been used in genetic disease studies, the development of NGS-appropriate methods is still in its infancy. Two novel approaches to exome sequencing studies of monogenic disease are described in this thesis.

1.5 Thesis Outline

This thesis will describe the development, assessment and application of two network-based methods to identify monogenic disease genes from exome sequencing data in the presence of genetic heterogeneity. To facilitate a clear and logical discussion of the work it is necessary to present the four main chapters (4, 5, 7 and 8) as separate "reports", each containing a short introduction, methods, results and discussion. The thesis is structured as follows.

Chapter 2 describes the data resources that are common to subsequent chapters, including the construction of interaction networks. Chapter 3 is a short chapter presenting results which motivate the subsequent work: interacting sets of genes which cause the same monogenic disease.

Chapters 4 and 5 report the development and performance assessment of the two methods. BioGranat-IG, discussed in chapter 4, seeks to identify small connected subnetworks of genes which harbour post-filtering variants in all (or most) exomes from of a cohort of unrelated affected individuals. Performance testing is undertaken using simulated data. In chapter 5, the HetRank approach is described, which is designed to overcome some

of the limitations of BioGranat-IG by using a variant-ranking approach. Simulated data are again used to assess the performance of HetRank and compare it to BioGranat-IG.

Two supporting methods which are used in both of the subsequent chapters, primarily for the interpretation of BioGranat-IG and HetRank results, are described fully in chapter 6.

In chapters 7 and 8 BioGranat-IG and HetRank, plus other simple network-based strategies, are applied to real exome sequencing studies. Findings will include putative disease-relevant pathways identified by the methods, and will enable a discussion of the utility of these tools in practice. Chapter 7 describes a study of Adams-Oliver syndrome, a monogenic disease with several known causal genes but demonstrated genetic heterogeneity. Chapter 8 analyses exome data from several families affected with Crohn's disease; while Crohn's disease is a complex disease the strong inheritance in these families suggests a prominent role for one or a small number of genes.

Finally, a short concluding discussion is presented in chapter 9.

2 Data Resources

2.1 Interaction Networks

2.1.1 Network Construction

The work presented in this thesis makes use of six interaction networks. Key properties of these networks are summarised in Table 2.1. Interaction data and networks were stored as plain text files; processing was performed using UNIX commands with simple Perl scripts and MySQL commands where required for more complex data manipulation.

- ***HuPPI2.*** This network is a protein interaction network (PIN) which integrates experimentally-verified interactions from six human protein-protein interaction (PPI) databases: BioGRID, MINT, BIND, DIP, IntAct and HPRD (Lehne and Schlitt 2009; Lehne 2011). Since multiple independent sources of evidence result in more reliable PPIs, the network includes only those interactions supported by two or more separate publications (Lehne 2011). Binary interactions are combined to give an unweighted and undirected network of 3,666 nodes (identified by Ensembl gene symbols) connected by 6,187 edges. The interaction data underlying the network was obtained in 2008; the network is not regularly updated and was obtained directly from Dr Thomas Schlitt on 19th October 2011.

- ***PINA.*** A more recent PIN is *PINA*, comprising interaction data from the publicly available PINA meta-database which integrates manually-curated PPIs from BioGRID, MINT, DIP, IntAct, HPRD and MIPS MPact (Cowley et al. 2012). Interaction data were downloaded on 20th December 2012 and comprised both self-interactions and binary interactions specified using UniProt gene names. Removal of the self-interactions gave an unweighted and undirected network of 14,380 nodes and 104,572 edges.*

* Note that the work described in chapter 5 is based on a version of this network for which self-interactions were not removed. The properties of this version of the network are similar (see Table 2.1) and the expected impact on results is negligible.

- ***PINAmin2.*** This network is a higher-confidence subnetwork of PINA, comprising only interactions supported by two or more independent publications. This results in reduced coverage of the genome relative to the full PINA network: 7,363 nodes are connected by 17,826 edges.[*]

- ***CPDBconf95.*** The ConsensusPathDB (CPDB) database combines literature-curated interactions and large-scale experimental data from 19 PPI databases (including all six databases underlying PINA) (Kamburov et al. 2013). Interactions are also assigned scores between 0 and 1 which combine three topology-based and three annotation-based measures of interaction confidence (Kamburov et al. 2012). These scores are most strongly influenced by two of the annotation-based measures: number of independent publications reporting an interaction, and similarity of Gene Ontology cellular component annotations (Kamburov et al. 2013 [Supplementary methods]).

  Interaction data were downloaded on 7th October 2013 (version 27) and comprised both self-interactions and binary interactions specified by UniProt IDs. These were converted to HUGO Gene Nomenclature Committee (HGNC) official gene symbols using a mapping file obtained from BioMart (Kasprzyk 2011). Removal of the self-interactions gave a weighted and undirected network, which was converted to an unweighted network using a "high-confidence" threshold of 0.95 (Kamburov et al. 2013 [Supplementary methods]) for the edge weights. The final network, *CPDBconf95*, consists of 6,136 nodes and 30,058 edges.

- ***COXPRES30.*** COXPRESdb is a gene co-expression database (Obayashi et al. 2013). For a given gene g, the other 19,000+ human genes included in the database are ranked according to the similarity between their expression profile and gene g's expression profile across a wide range of microarray samples. The mutual rank between gene g_1 and gene g_2 is the (geometric) average of g_1's rank among genes similar to g_2 and g_2's rank among genes similar to g_1; mutual rank was found to be a better measure of relatedness than directly quantifying expression profile similarity (Obayashi et al. 2008).

  Mutual ranks for all gene pairs, encoded by Entrez IDs, were downloaded on 9th April 2014 (version c4.1), and converted to HGNC official gene symbols

[*] Note that the work described in chapter 5 is based on a version of this network for which self-interactions were not removed. The properties of this version of the network are similar (see Table 2.1) and the expected impact on results is negligible.

using the mapping file provided by COXPRESdb. An unweighted and undirected network, *COXPRES30*, was constructed where edges connect pairs of genes having mutual rank <30. This corresponds to medium- and high-similarity co-expression pairs (Obayashi et al. 2008) and gives a network of 18,454 nodes and 128,688 edges.

- ***Multinet.*** This is a "unified" network in which edges can represent one of several different types of relationship between genes. Interactions comprise PPIs (extracted from BioGRID), as well as genetic, regulatory, metabolic, signalling and phosphorylation (kinase-substrate) relationships (Khurana et al. 2013). The authors suggest that different interaction types are needed to comprehensively model how genes and their protein products collaborate to form a functioning system, and find a very low level of redundancy between interaction types (Khurana et al. 2013). Interaction data, comprising binary interactions specified by gene symbols, were downloaded on 13th February 2014; these resulted in an unweighted and undirected network of 14,445 nodes and 109,598 edges.

For all six networks, the majority of nodes are connected in a single large component (see Table 2.1).

Figure 2.1 shows plots of the degree distributions for all six networks. The networks show a broadly scale-free topology, meaning that the degree distributions follow a power law. However, the distribution for the COXPRES30 network appears to differ from those of the other networks due to a poorer fit at lower degrees, suggesting a slightly different topology with a relatively higher proportion of medium-degree nodes. The properties of this network in Table 2.1 also suggest an atypical topology; notably it has a lower maximum degree and higher large component average path length than other networks of comparable size. These topological differences most likely reflect the different data types underlying the networks: while the other networks mainly comprise physical interactions between proteins, COXPRES30 is based on pairwise expression profile similarity (and the mutual rank measure prevents nodes from having extremely high degree).

2.1.2 Hub Removal

It has been shown that removing network nodes of high degree (hubs) can lead to improved identification of functional modules (Liu et al. 2011) and disease-associated subnetworks (Lehne 2011) within an interaction network.

HuPPI2_d25, a subnetwork of HuPPI2 derived by removing nodes with degree 25 or greater, was obtained directly from Dr Thomas Schlitt on 19th October 2011.

Table 2.1 – Key properties of interaction networks

Network	Interaction type	Nodes	Edges	Average Degree	Maximum Degree	Nodes in Large Component	Large Component Average Path Length
HuPPI2	PPI	3,666	6,187	3.38	108	3,027	5.88
PINA	PPI	14,380	104,572	14.54	7,804	14,326	2.91
(with self-interactions retained)		*14,434*	*105,801*	*14.66*	*7,804*	*14,326*	*2.91*
PINAmin2	PPI	7,363	17,826	4.84	4,778	7,128	2.84
(with self-interactions retained)		*7,417*	*18,092*	*4.88*	*4,778*	*7,128*	*2.84*
CPDBconf95	PPI	6,136	30,058	9.80	850	5,768	4.23
COXPRES30	Co-expression	18,454	128,688	13.95	282	18,360	5.11
Multinet	Integrated	14,445	109,598	15.17	1,496	14,399	3.39

Table 2.2 – Key properties of interaction networks following hub removal

Network	Interaction type	Nodes	Edges	Average Degree	Maximum Degree	Nodes in Large Component	Large Component Average Path Length
HuPPI2_d25	PPI	3,433	4,630	2.70	23	2,659	7.53
PINA_d50	PPI	10,375	29,803	5.75	41	10,141	5.11
(with self-interactions retained)		*10,481*	*30,950*	*5.91*	*43*	*10,141*	*5.11*
PINAmin2_d50	PPI	4,565	9,561	4.19	41	4,022	5.86
(with self-interactions retained)		*4,647*	*9,820*	*4.23*	*41*	*4,022*	*5.86*
CPDBconf95_d50	PPI	5,548	16,247	5.86	47	5,111	5.48
COXPRES30_d50	Co-expression	17,001	90,183	10.61	49	16,860	5.60
Multinet_d50	Integrated	8,199	23,605	5.76	43	7,955	5.39

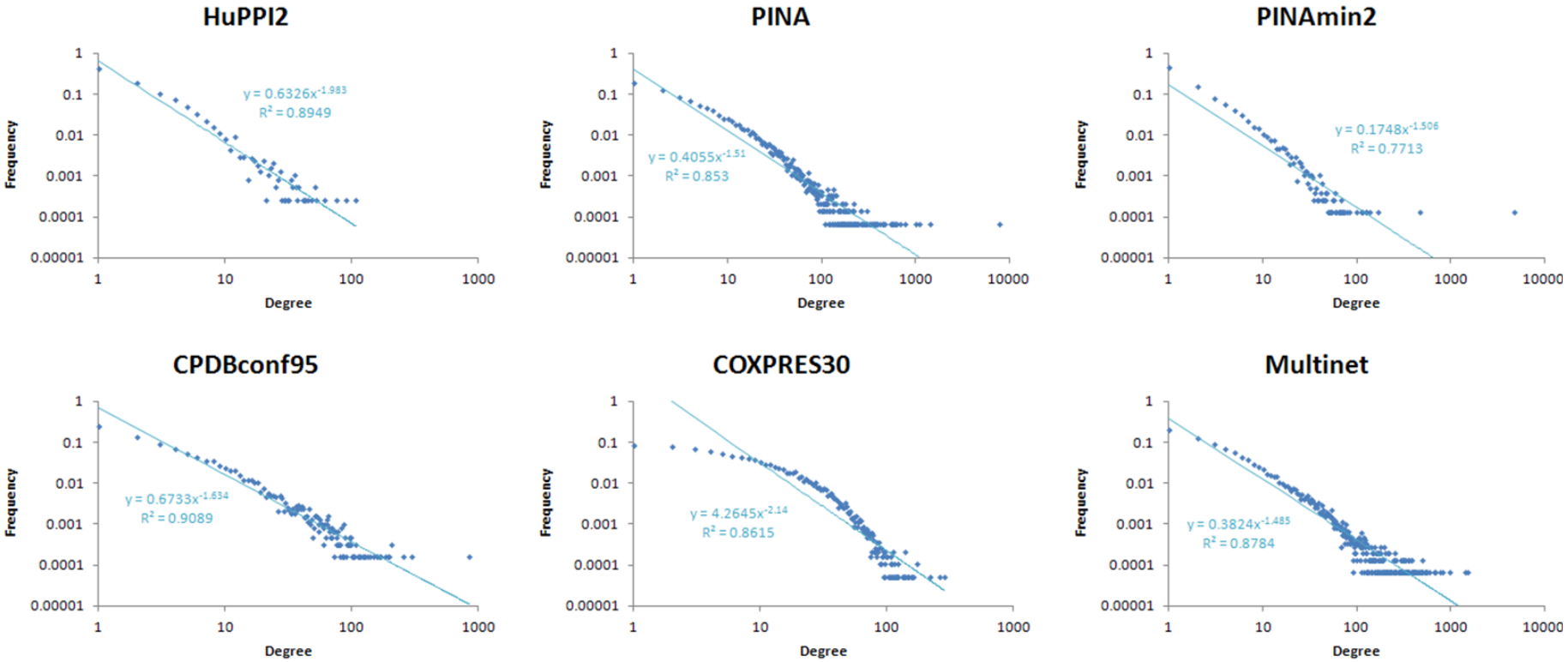


Figure 2.1 – Degree distributions of interaction networks
For each network, node degrees are plotted against the fraction of nodes in the network of that degree. Regression is performed assuming a power-law trend line. Values are plotted against a logarithmic scale on both axes.

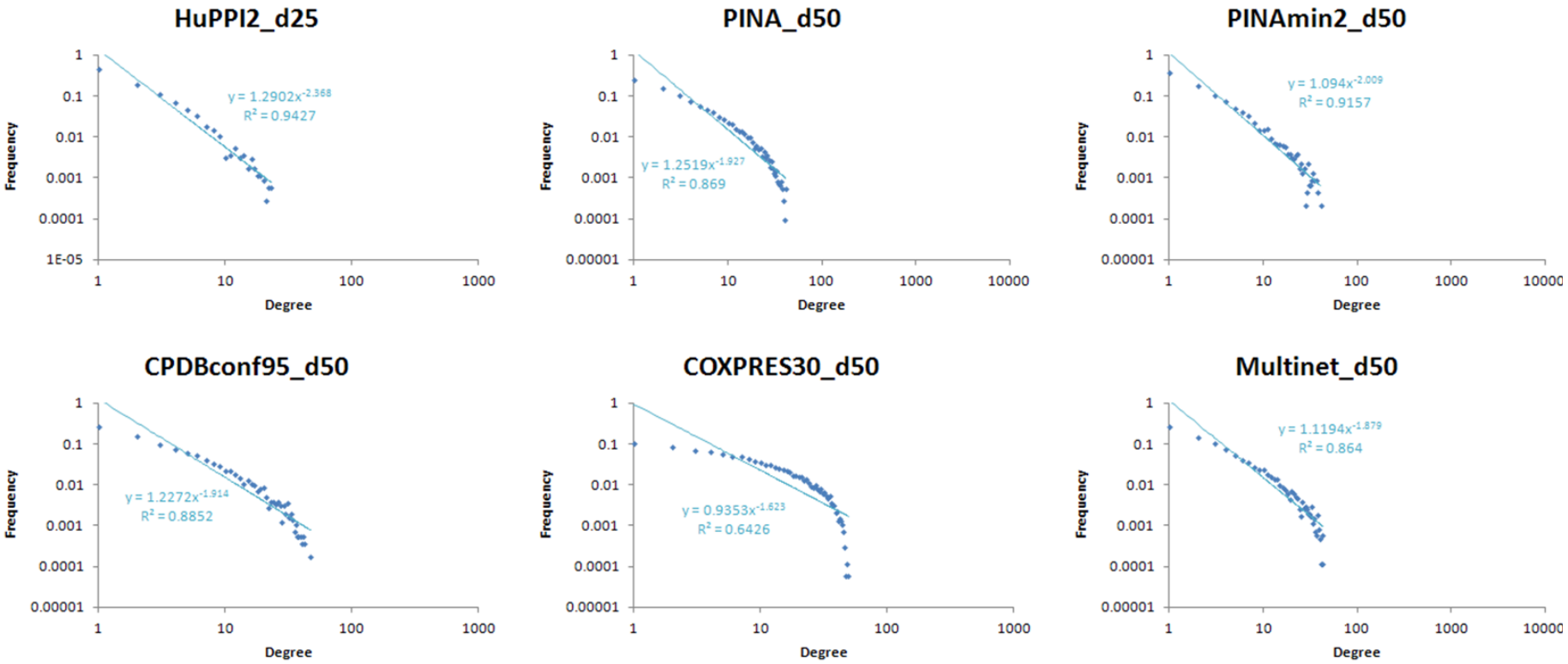


Figure 2.2 – Degree distributions of interaction networks following hub removal
For each network, node degrees are plotted against the fraction of nodes in the network of that degree. Regression is performed assuming a power-law trend line. Values are plotted against a logarithmic scale on both axes.

For all other networks, hub-free versions with the suffix "*_d50*" were derived by removing nodes with degree 50 or greater.[*] Hub removal was performed in R using the igraph library for network analysis (Csardi and Nepusz 2006).

Table 2.2 gives the key properties of the networks following hub removal. In all cases the hub-free networks have a lower average degree and higher large component average path length than the corresponding original networks. Figure 2.2 shows the degree distributions for these networks, which are truncated relative to the full degree distributions in Figure 2.1.

2.1.3 Network Agreement

Given that many of these networks share common underlying data sources (and in particular that the PINs are intended to describe the same underlying interactome of PPIs), some overlap between the networks is expected. Table 2.3 shows the proportion of nodes and edges (that is, genes and interactions) that are shared between the networks.

Table 2.3a gives the overlap between the full networks before hub removal. As expected, all of PINAmin2's nodes and edges are contained in PINA because it is a subnetwork. PINA also includes the majority of nodes from the other high-confidence PINs, HuPPI2 and CPDBconf95 (over 97% in each case), and covers 90.6% and 62.2% of their edges respectively. The agreement between PINA and Multinet is more limited, with 79.1% of Multinet's nodes and 35.3% of its edges present in PINA (and roughly similar proportions of PINA's nodes and edges present in Multinet). Strikingly, while COXPRES30 covers the majority of nodes found in each of the other networks, there is very little correspondence between edges. The interactions in COXPRES30, where an edge signifies that genes have highly similar expression profiles, largely therefore describe a distinct type of functional relationship compared to the mainly physical interactions described by the other networks.

Table 2.3b gives the overlap between the hub-free networks. Although there are a few exceptions, what we largely see is a fall in the degree of sharing of both nodes and edges compared to the full networks. This is due to the fact that hub removal affects different genes in each network. However, we see broadly the same patterns overall: most of the nodes from the smaller high-confidence networks are seen in the bigger networks, but fewer of the edges; the two bigger networks of mainly physical interactions have a considerable overlap of edges, with 45.7% of Multinet_d50's present in PINA_d50 and 36.2% of

[*] A higher threshold is used for the networks other than HuPPI2 because they are larger and more densely connected. However, the thresholds of 25 and 50 are somewhat arbitrary; a more rigorous approach to hub removal is proposed in section 9.2.3 (in the Future Work section of the Concluding Discussion).

Table 2.3 – Agreement of nodes and edges between networks
Grey shading denotes same network; red shading denotes <10% overlap; yellow shading denotes >50% overlap; green shading denotes >90% overlap.

(A) Hub-retained networks

Nodes from:	**In network:**					
	HuPPI2	PINA	PINAmin2	CPDBconf95	COXPRES30	Multinet
HuPPI2	100.0%	97.0%	84.4%	74.6%	93.8%	96.0%
PINA	24.7%	100.0%	51.2%	41.7%	85.4%	79.4%
PINAmin2	42.0%	100.0%	100.0%	59.9%	92.6%	91.2%
CPDBconf95	44.5%	97.8%	71.9%	100.0%	93.6%	91.9%
COXPRES30	18.6%	66.6%	36.9%	31.1%	100.0%	66.7%
Multinet	24.4%	79.1%	46.5%	39.0%	85.2%	100.0%

Edges from:	**In network:**					
	HuPPI2	PINA	PINAmin2	CPDBconf95	COXPRES30	Multinet
HuPPI2	100.0%	90.6%	60.8%	46.2%	3.4%	68.8%
PINA	5.4%	100.0%	17.1%	17.9%	1.0%	37.0%
PINAmin2	21.1%	100.0%	100.0%	35.9%	2.2%	56.3%
CPDBconf95	9.5%	62.2%	21.3%	100.0%	5.1%	38.2%
COXPRES30	0.2%	0.8%	0.3%	1.2%	100.0%	0.6%
Multinet	3.9%	35.3%	9.2%	10.5%	0.7%	100.0%

(B) Hub-free networks

Nodes from:	**In network:**					
	HuPPI2 _d25	PINA _d50	PINAmin2 _d50	CPDBconf95 _d50	COXPRES30 _d50	Multinet _d50
HuPPI2_d25	100.0%	81.3%	79.2%	70.4%	86.3%	78.7%
PINA_d50	26.9%	100.0%	37.2%	45.3%	81.4%	66.7%
PINAmin2_d50	59.6%	84.5%	100.0%	68.2%	86.2%	78.3%
CPDBconf95_d50	43.6%	84.6%	56.1%	100.0%	86.0%	73.3%
COXPRES30_d50	17.4%	49.7%	23.1%	28.1%	100.0%	41.1%
Multinet_d50	33.0%	84.4%	43.6%	49.6%	85.2%	100.0%

Edges from:	**In network:**					
	HuPPI2 _d25	PINA _d50	PINAmin2 _d50	CPDBconf95 _d50	COXPRES30 _d50	Multinet _d50
HuPPI2_d25	100.0%	43.6%	56.8%	36.5%	3.1%	36.6%
PINA_d50	6.8%	100.0%	12.5%	20.4%	1.2%	36.2%
PINAmin2_d50	27.5%	39.0%	100.0%	35.1%	2.6%	35.2%
CPDBconf95_d50	10.4%	37.5%	20.6%	100.0%	2.7%	23.6%
COXPRES30_d50	0.2%	0.4%	0.3%	0.5%	100.0%	0.3%
Multinet_d50	7.2%	45.7%	14.3%	16.3%	1.3%	100.0%

PINA_d50's present in Multinet_d50; and again, COXPRES30_d50 has many nodes in common with the other networks but very few edges.

2.2 BioGranat Software

BioGranat ("Molecular Biology Graph Visualisation and Analysis Tool") is a software tool for the analysis and visualisation of biological interaction networks (Mendig et al. 2009). It is implemented using the Java OSGi framework and is freely available from www.biogranat.org. BioGranat was developed by Dr Thomas Schlitt at King's College London (KCL) in collaboration with Prof. Volker Ahlers and Prof. Frauke Sprengel at the University of Applied Sciences and Arts in Hannover, Germany.

BioGranat provides a user interface and basic network analysis functionality, allowing customised tools for specific analyses to be developed as BioGranat bundles. BioGranat-IG (described in chapter 4 and applied in chapters 7 and 8) and Region Growing Analysis (discussed in chapter 6 and used extensively in chapters 7 and 8) are both BioGranat bundles.

2.3 Exome Sequencing and Annotation

All whole exome sequence data (including disease cases, healthy controls and exomes used to generate test data) were obtained as annotated output files from the exome sequencing pipeline employed by the KCL rare disease programme, overseen by Dr Michael Simpson.

The majority of whole exome samples were sequenced at the NIHR Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and KCL. For these exomes, whole-exome capture was performed using the Agilent Sure Select XT Human All Exome Kit. Enriched DNA fragments were sequenced using an Illumina HiSeq 2000 instrument. The exomes of 45 Crohn's disease cases (analysed in chapter 8) were sequenced at the Institute of Clinical Molecular Biology at the University of Kiel, Germany. For these exomes, whole-exome capture was performed using the Nextera Rapid Capture Expanded Exome kit and enriched DNA fragments were also sequenced using an Illumina HiSeq 2000 instrument.

For all sequenced exomes, paired-end reads were aligned to the UCSC Genome Browser's hg19 reference sequence (Karolchik et al. 2014) using Novoalign (v.2.07.17; Novocraft Technologies). SNVs and small indels were called using SAMtools (v.0.1.18; Li et al. 2009) and annotated with gene and transcript identifiers from RefSeq (Pruitt et al.

2014) using ANNOVAR (2013Feb21 version; Wang et al. 2010c). Variants were also annotated with reference to 1000 Genomes Project and Exome Variant Server (EVS) data (1000 Genomes Project Consortium 2010; NHLBI Exome Sequencing Project 2014) and ~850 exomes sequenced in-house (referred to in subsequent chapters as the *in-house exome database*).

3 Motivation: Interacting Genes Cause the Same Monogenic Disease

3.1 Introduction

It has previously been established that pairs of genes causing the same disorder are significantly likely to be connected in a protein interaction network (PIN). Feldman *et al.* showed that significantly many genes causing the same (mainly monogenic) diseases form connected clusters in a network (Feldman et al. 2008). Likewise, Goh *et al.* found significant overlap between the connections in a PIN and the connections in a network formed by linking two genes if they are associated with the same monogenic or complex disease (Goh et al. 2007). Using a more general notion of phenotype, Van Driel *et al*. demonstrated that a text-based measure of similarity between disease phenotypes is correlated with the connectedness of causal genes in a PIN (van Driel et al. 2006).

The use of interaction networks to address genetic (locus) heterogeneity is central to this thesis. To motivate our approach, and to generate working examples on which to base simulated test data, we therefore undertook an independent investigation of the extent to which genes causing the same monogenic disease interact in PINs.

For two PINs (PINA and the high-confidence subnetwork PINAmin2) we identified *disease subnetworks*, defined as directly-connected sets of two or more genes that are designated in OMIM (Amberger et al. 2009) as being causal for the same monogenic disease (disregarding disease sub-types). By comparison against randomly-permuted networks we found that significantly many diseases display locus heterogeneity that can be modelled by disease subnetworks.

3.2 Methods

Analysis was performed separately using the PINA and PINAmin2 networks described in section 2.1.1; all data analysis was performed in R using the igraph library for network analysis (Csardi and Nepusz 2006).

Disease-gene mappings were obtained from OMIM's Morbid Map (downloaded 20th March 2013) (Amberger et al. 2009). Unconfirmed mappings and mappings which either involve non-disease phenotypes or where the gene is not directly causal (such as genes

which contribute to susceptibility to a multifactorial disorder or infection) were excluded, leaving 4,956 monogenic disease mappings.

OMIM disease terms were replaced with generalised disease terms by removing disease "type" or "group" names, and any words consisting entirely of numbers. This resulted in 3,193 generalised disease terms in total. 541 diseases displayed locus heterogeneity by mapping to more than one causal gene.

In each of the two PINs, OMIM disease subnetworks were found by considering one generalised disease term at a time and identifying direct interactions between causal genes.

To test the null hypothesis that the disease subnetworks found could arise by chance due to the number of disease genes mapping into each network, disease subnetworks were also identified in 10,000 randomisations of each PIN. To counter the potential bias in curated interaction networks towards well-studied disease-causing genes, the degree-constrained network permutation approach described in (Lehne 2011) was used. Briefly, node labels are preferentially swapped with nodes of similar degree. For node *g* this is achieved by listing all other network nodes in increasing order of degree difference relative to *g* (with nodes of equal degree ordered uniformly randomly); a one-tailed normal distribution centred at the top of this list is then used to select a node and labels are swapped. The default standard deviation of 5.0 nodes was used.

3.3 Results and Discussion

172 connected subnetworks were found in the PINA network, and 84 in the PINAmin2 network, each of which is causal for a single disease. Table 3.1 summarises these findings, and a full list of disease subnetworks is provided in Appendix A.

Table 3.1 also gives the average number of disease subnetworks found in 10,000 permutations of each network, the distributions of which are illustrated in Figure 3.1. In each case the observed number of disease subnetworks is highly significant ($p < 10^{-4}$), enabling us to conclude that these disease subnetworks could not arise by chance.

The methods used here are similar to those employed by Feldman *et al.*, and these results corroborate their findings, although the number of disease subnetworks found exceeds the 30 specific to monogenic diseases that they reported in 2008 (Feldman et al. 2008). As expected our results support the broader assertion that interacting genes are more likely to have similar phenotypic consequences (van Driel et al. 2006; Goh et al. 2007). Even allowing for the possibility that our degree-constrained permutation approach insufficiently overcomes the potential knowledge bias in curated interaction networks, this still makes a compelling case that genes causing similar or identical clinical phenotypes

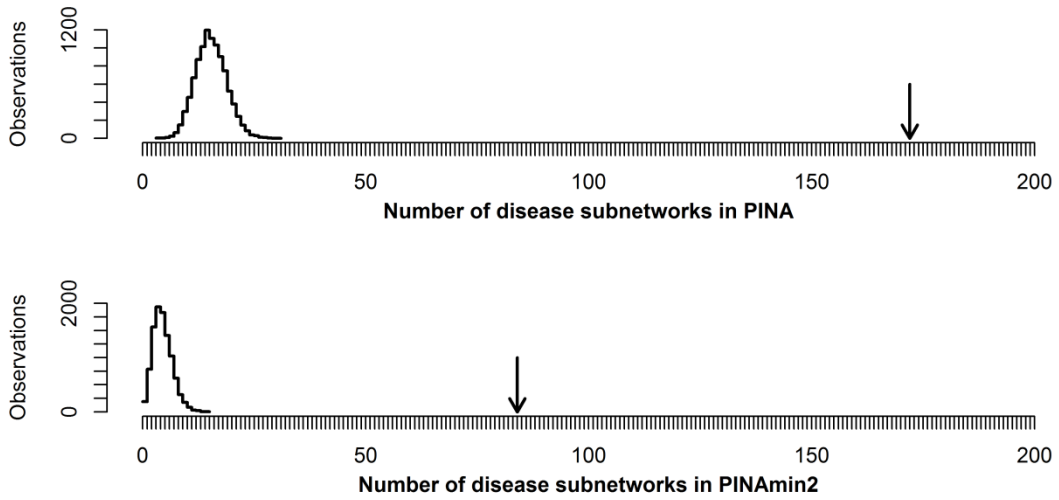


Figure 3.1 – Interactions between genes involved in the same disease occur frequently
Frequency plots showing the number of disease subnetworks (connected sets of two or more genes causing the same disease) identified in 10,000 random permutations of the PINA and PINAmin2 networks. Arrows indicate number of disease networks found in the original networks.

Table 3.1 – OMIM disease subnetworks
Observed = number of disease subnetworks of given size induced by a single disease term in the original network; Permutation Average = average number of disease subnetworks of given size induced by a single disease term across 10,000 randomly permuted networks (mean ± standard deviation).

Size of disease subnetwork	Number of disease subnetworks			
	PINA network		PINAmin2 network	
	Observed	Permutation average	Observed	Permutation average
2	124	12.54 ± 3.35	61	3.46 ± 1.93
3	27	1.65 ± 1.25	11	0.44 ± 0.73
4	11	0.40 ± 0.60	7	0.06 ± 0.25
5+	10	0.28 ± 0.49	5	0.07 ± 0.26
Total	172	14.87 ± 3.46	84	4.03 ± 2.13

frequently interact and thus it makes sense to use interaction networks as a means to identify new sources of genetic heterogeneity, particularly given that high-throughput methods are continually improving network coverage (Yu et al. 2011).

In addition to re-iterating this important observation, the disease subnetworks identified here can be used to generate realistic test data with which to assess the performance of the methods developed in this thesis. In chapter 5 randomised exome data will be "spiked" with disease-causing mutations in genes drawn from disease subnetworks. The ability to recover these disease subnetworks will be an important measure of the performance of the BioGranat-IG and HetRank methods, and of the extent to which they improve upon simple intersection filtering.

It could be argued that the use of generalised disease terms, which disregard disease "type" or "group", will conflate sub-types of a disease that could display phenotypic differences. While in some cases this may be true, many disease sub-types in OMIM represent different molecular bases of a given disorder. Moreover, in practice – due to limitations of available data – whole exome sequencing studies of rare diseases often seek to identify a causal gene using exomes from a group of affected individuals with the same clinical diagnosis but some phenotypic variability. Therefore the disease subnetworks identified here provide a reasonable basis on which to test analysis methods designed to address genetic heterogeneity in such studies.

4 Development of BioGranat-IG Analysis Tool

4.1 Introduction

As previously discussed, improvements in sequencing technology have led to considerable successes in the identification of disease-causing sequence variants for several rare monogenic diseases. Typically intersection filtering is used, in which a series of filtering steps are applied to whole exome sequence variants from several unrelated affected individuals (e.g. Ng et al. 2010a; Simpson et al. 2011). However, depending on the disease, one might not observe any genes in the intersection of the filtered lists. One phenomenon that can cause this is genetic heterogeneity, where one phenotypic outcome results from any one of a number of possible mutations, possibly in different loci (McClellan and King 2010). An example would be two genes that are functionally related through their protein products, both playing a critical role in some cellular function and such that a mutation in either gene leads to the failure of that function.

One possible way to deal with this problem would be to look for the smallest set of genes such that all individuals carry a post-filtering sequence variant in the set. This frames the problem as one of minimal set cover (MSC; Cormen 2001) and will not guarantee any biological meaning or functional relatedness for the set of genes found.

Instead, in this chapter, we present BioGranat-IG ("BioGranat Individuals-Grouping"), a tool that uses the additional structure found in interaction networks to analyse sequence data for multiple individuals and suggest possible sources of genetic heterogeneity.

Interaction networks are ideal for this purpose because they connect genes for which a functional relationship is known – or predicted – to exist. Several tools, such as GeneMANIA (Warde-Farley et al. 2010) and STRING (Franceschini et al. 2013), exist that use interaction networks to suggest additional functionally related genes for a given input gene list. Conversely, our tool will use the networks to focus on which of the input genes are likely to be causative and provide a relatively short list of candidate genes for follow-up study.

The premise behind BioGranat-IG is that if a disease has an underlying mechanism of locus heterogeneity, the genes involved may be functionally related and therefore closely connected in an interaction network; if not directly connected, they might be interacting with one or more common neighbours. To identify disease gene candidates, given a number of individuals and for each individual a list of genes with sequence variants (filtered in the

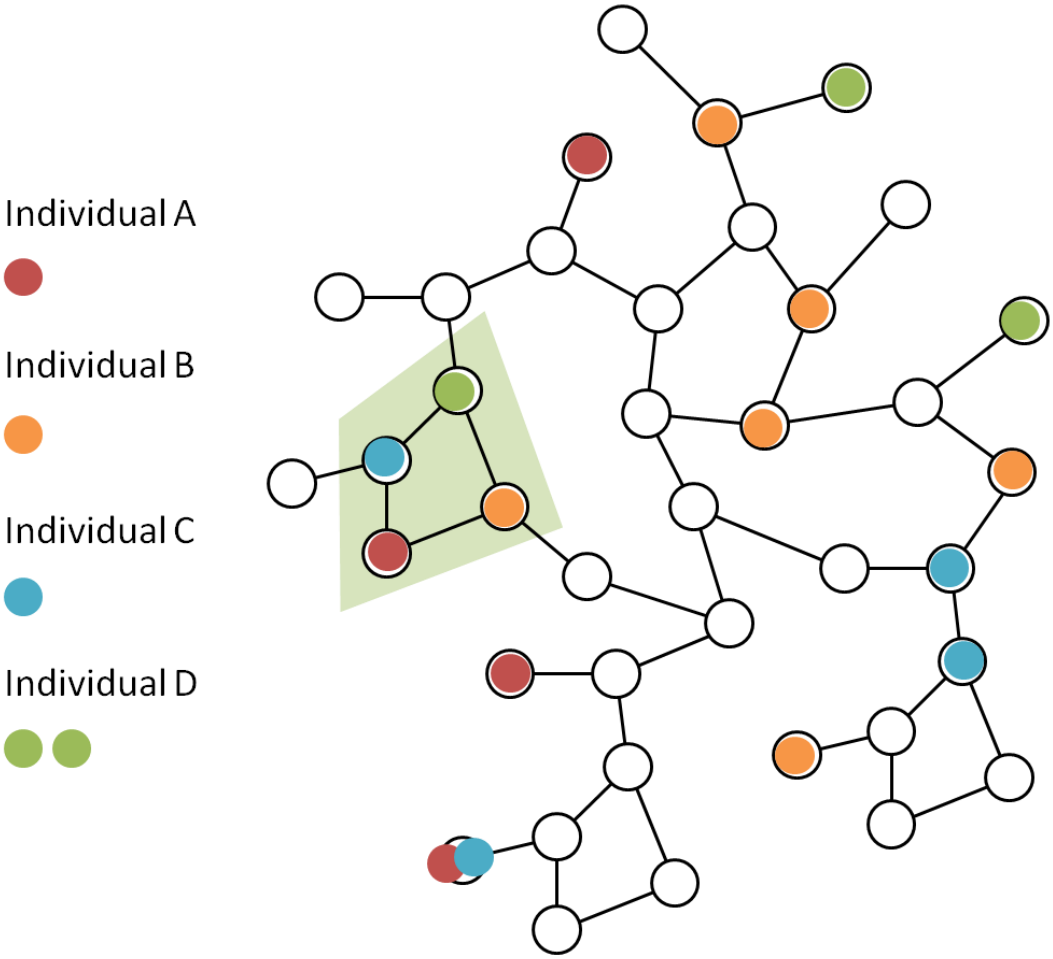


Figure 4.1 – BioGranat-IG strategy for gene identification
For individuals A-D, genes containing variants that have not been removed during the filtering step are represented as red, yellow, blue and green tokens respectively. These tokens are mapped to the corresponding nodes in a gene network. Not every token can be mapped to the network because gene networks typically do not cover the entire genome. The smallest connected subnetwork to cover all individuals is marked by the green diamond.

usual manner based on novelty or variant frequency, functional consequence etc.), we mark each node in our network with the individuals who carry a mutant version of that gene. Subsequently, we seek the smallest connected subnetwork marked with all individuals (see Figure 4.1).

Our network thus consists of "marked" nodes, representing genes where a sequence variant was observed in some individual, and "empty" nodes, where no variant was observed for any individual. As an additional constraint on the subnetwork we seek, we require that each marked node is connected to another marked node via at most one empty node. That is, the subnetwork is allowed to incorporate "jumps" of one empty node, but not more (see Figure 4.2). The rationale underlying this is that interaction networks tend to exhibit the small world phenomenon, meaning that the average path length between any two nodes is

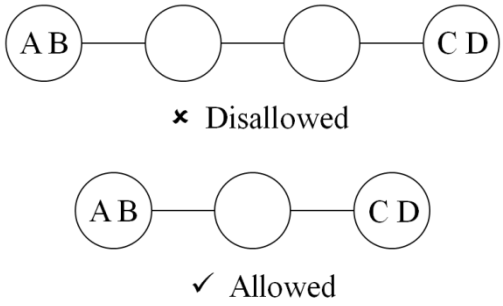


Figure 4.2 – Allowed jumps in BioGranat-IG
Subnetworks returned by BioGranat-IG can connect marked nodes via jumps of one empty node (a gene in which no individuals have a mutation) but not more. These examples use individuals labelled A–D.

short, typically four to six edges (Xu et al. 2011). We are therefore interested only in localised connections, and allowing too many empty nodes into our subnetwork could reduce the chance of it having any biological meaning. This constraint allows a computationally efficient implementation, making it possible to run BioGranat-IG on large sets of permuted data to establish statistical significance.

This chapter will present the methods implemented by BioGranat-IG and demonstrate the validity of the approach using simulated whole exome sequencing output. The performance of the tool under various conditions is analysed, before considering the outstanding challenges of the approach.

To our knowledge, BioGranat-IG is the first tool developed to tackle the problem of finding disease-causing genes from whole exome sequence data for monogenic diseases in the presence of locus heterogeneity. As discussed at length in chapter 1 (section 1.4.6), finding dysregulated subnetworks is an intensively studied problem (Lehne and Schlitt 2012; Staiger et al. 2012), but few approaches consider data for individuals separately. Rather, most approaches work with summary statistics. Notably, a method by Dao *et al.* (Dao et al. 2010), DEGAS (Ulitsky et al. 2010) and KeyPathwayMiner (Alcaraz et al. 2012) use differential expression data for individuals to find subnetworks containing genes differentially expressed in patients versus controls. While for differential gene-expression data one usually expects to find clusters of co-expressed functionally related genes, our problem differs because we expect all individuals to carry only a limited number (probably less than tens) of disease-causing genes hidden among a large number (hundreds) of variants not related to the disease of interest. Therefore, the problem addressed by KeyPathwayMiner and DEGAS is similar to the problem we address here in general, but there are important differences in the detail that have an impact on the algorithm design.

The DAPPLE tool (Rossin et al. 2011) prioritises genes in genomic regions associated with a disease using a network-based approach conceptually related to BioGranat-IG. However, DAPPLE is designed to improve understanding of disease-associated loci, and is not readily applicable to the sequencing problem we describe.

The work presented in this chapter has been published (Dand et al. 2013) and is reproduced here by permission of Oxford University Press. Co-author contributions are as follows: Prof. Volker Ahlers and Prof. Frauke Sprengel at the University of Applied Sciences and Arts in Hannover, Germany, developed BioGranat in collaboration with Dr Thomas Schlitt at King's College London; Dr Schlitt also initiated and obtained funding for network method development projects (including my PhD studentship), conceived the analysis strategy and oversaw analysis of results. The algorithm design and implementation, analysis, and interpretation of results presented in this chapter were performed by Nick Dand.

4.2 Methods

BioGranat-IG has been developed as a BioGranat bundle (see chapter 2) and is available from www.biogranat.org.

In graph-theoretic terms, the problem we face can be expressed as follows. Let S_n be the set of n elements {1, 2, ..., n}. The power set $P(S_n)$ of S_n is the set of all possible subsets of S_n, including the empty set. Then, given a graph G with nodes $V(G)$, and for each g in $V(G)$ a mapping $f(g)$ into $P(S_n)$, we wish to find the smallest connected subgraph G' of G such that:

$$\bigcup_{g \in V(G')} f(g) = S_n$$

Here, smallest is taken to mean least number of nodes. It is possible that no such subgraph exists, in which case we seek the smallest connected subgraph such that:

$$\left| \bigcup_{g \in V(G')} f(g) \right| = m,$$

where m is the maximum number of elements of S_n that are mapped to by the nodes of a single connected component of G. (Note that in this chapter, we refer more loosely to seeking small subnetworks "containing" all individuals).

The problem is an example of the minimal connected set cover problem (MCSC) (Cerdeira and Pinto 2005; Zhang et al. 2009), which is NP-hard because it is a generalisation of the MSC problem (Karp 1972). Several authors have published approximation algorithms for MCSC in recent years (Ren and Zhao 2011; Elbassioni et al. 2012). However, for BioGranat-IG, we have developed a new method because we want to collect not just the size of the optimal subnetwork, but all examples of optimal and near-optimal subnetworks, up to a user-specified size. BioGranat-IG cannot determine which subnetworks will be of most interest to the user, and so must output them all.

The following sections present the methods used by BioGranat-IG to find near-optimal small subnetworks containing variants for maximum individuals.

4.2.1 Network Pre-processing

BioGranat-IG works by marking lists of genes for multiple individuals onto an interaction network. There are >20,000 human genes but the function and expression of many genes is only poorly understood. Therefore, currently available networks are all incomplete. Nevertheless, they contain thousands of nodes and edges. To speed computation, the network is pre-processed (see Figure 4.3).

- Because jumps of more than one empty node are not allowed, all edges with an empty node at both ends are removed. From this point on, any neighbour of an empty node must be a marked node.
- All empty nodes of degree zero or one are removed.
- Where two empty nodes are connected to the same set of neighbours, one of the nodes can be removed from the network (and stored to provide an alternative result should the kept node turn out to form part of a minimal subnetwork).
- Any empty node whose neighbours form a clique (a complete subnetwork) is removed from the network. Such a node will never be called on to link two marked nodes.

4.2.2 Triplet and Quadruplet Search

Before resorting to heuristic methods, which cannot guarantee that all minimal subnetworks are returned, BioGranat-IG performs two searches (triplet search and quadruplet search), which together comprise an exhaustive search of all subnetworks of up to four nodes. If a subnetwork is found that covers all individuals, there is no need to perform any subsequent searches.

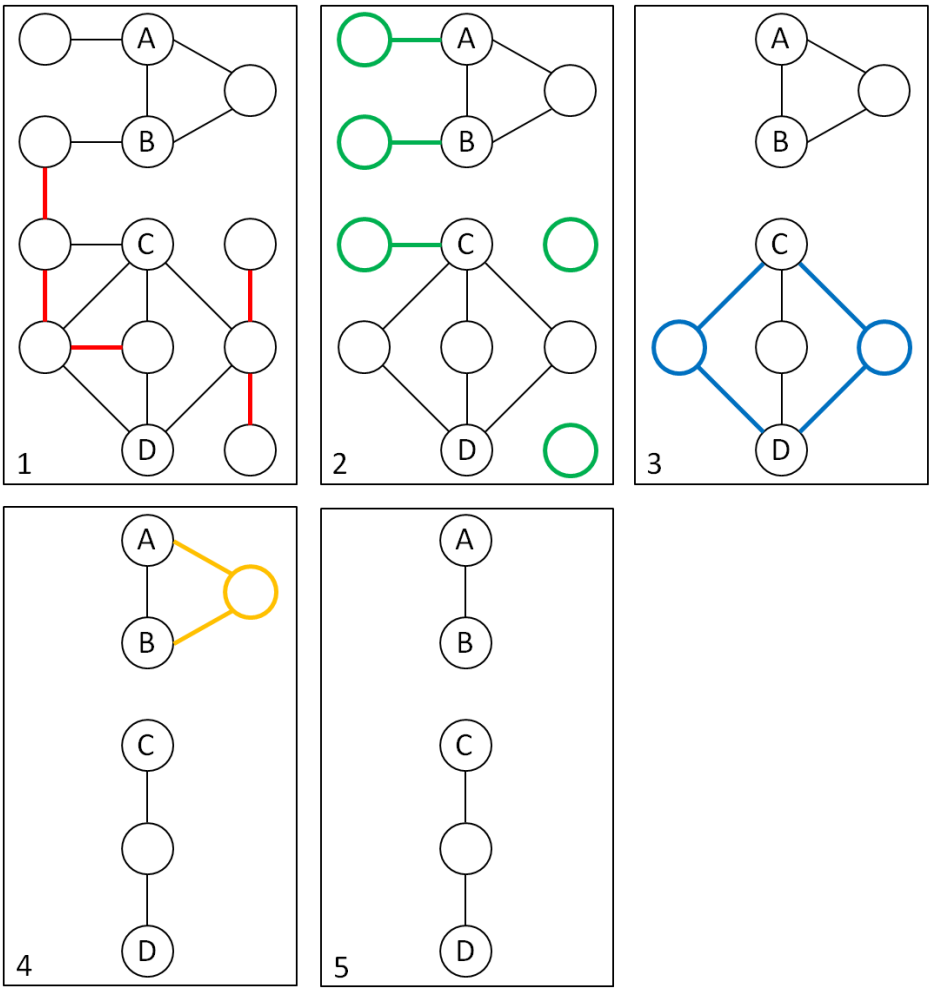


Figure 4.3 – Network pre-processing
Panel 1 shows a graph labelled with individuals A-D. First, edges connecting two empty nodes (marked in red) are removed, giving the graph in panel 2. Next, nodes of degree zero or one (marked in green) are removed, giving the graph in panel 3. Next, alternative connecting empty nodes (marked in blue) are removed, giving the graph in panel 4. These alternative nodes are stored because they could potentially be needed later to generate alternative optimal subnetworks. Finally, empty nodes whose neighbours form a clique (marked in yellow) are removed, giving the final pre-processed graph in panel 5.

The triplet search identifies all candidate subnetworks of up to three nodes, using the fact that for three nodes to be connected there must be at least one path of length two connecting them. We call the subnetwork induced by such a path a triplet. For each node in the network (whether marked or empty) we first check whether this node alone contains all individuals, and then identify the neighbouring nodes. All pairs formed from the original node plus one marked neighbour, and triplets formed from the original node plus two marked neighbours are examined. At this point, if a subnetwork is found containing all individuals, there is no need to continue with the quadruplet search.

The quadruplet search builds on the triplet search. Using the constraint that only one empty node can be jumped at a time, we know that for a subnetwork of four nodes to minimally cover all individuals it must contain at least one triplet with two individuals. Using this set of triplets as a starting point, quadruplets are thus constructed through the addition of any neighbouring nodes that confer additional individuals.

4.2.3 Minimum Distance Search

If the triplet and quadruplet searches fail to find a subnetwork covering all individuals, BioGranat-IG will perform heuristic searches based on a minimum distance approach (described here) and a multi-minimum distance approach (described in the next section).

The minimum distance search uses a greedy approach to build subnetworks starting from a single node. The selection function used to determine the most valuable neighbour to add to the subnetwork at each step is the sum of the minimum distances (length of shortest path) from each neighbour to all individuals not already covered by the subnetwork.

This approach requires that for every node in the network, the minimum distance to each individual is calculated. For node g, the distances $\{d_1(g), ..., d_n(g)\}$ represent the minimum distance from g to any node that contains individual 1, ..., n, respectively. Distances are calculated using a multi-source breadth-first search approach, as described by the following pseudocode:

1: For each individual, i
2: For each node g in the component
3: If g is marked with i, set $d_i(g) = 0$, and add g to the queue
4: Else set $d_i(g) = \infty$
5: While the queue is not empty
6: Take node g from the queue, and for all neighbours g' of g
7: If $d_i(g') = \infty$, set $d_i(g') = d_i(g) + 1$, and add g' to the queue

Because the pre-processed network may consist of several components, infinite distances can remain.

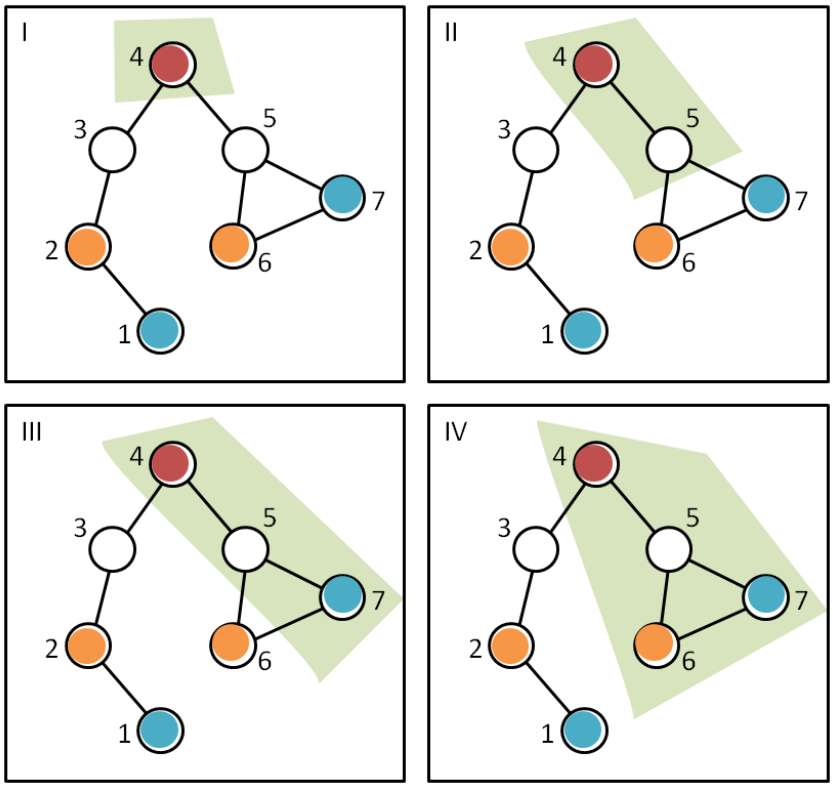


Figure 4.4 – Minimum distance search
A network of genes labelled 1-7 has been marked with individuals *R* (red tokens), *B* (blue tokens) and *Y* (yellow tokens). Panels I-IV illustrate how a subnetwork is built up (indicated by green shading), starting from gene 4. Let $d_X(n)$ indicate the minimum distance from node *n* to any node marked with individual *X*. Panel I: the subnetwork contains node 4 and covers individual *R*. The neighbouring nodes are 3 and 5. Since $d_B(3) + d_Y(3) = 2 + 1 = 3$ and $d_B(5) + d_Y(5) = 1 + 1 = 2$, node 5 is added next. Panel II: the subnetwork contains nodes 4 and 5 and covers individual *R*. The neighbouring nodes are 3, 6 and 7. Since $d_B(3) + d_Y(3) = 3$, as before, $d_B(6) + d_Y(6) = 1 + 0 = 1$ and $d_B(7) + d_Y(7) = 0 + 1 = 1$, either node 6 or 7 can be added next. Panel III: The subnetwork is 4-5-7 and covers individuals *R* and *B*. The neighbouring nodes are 3 and 6. Since $d_Y(3) = 1$ and $d_Y(6) = 0$, node 6 is added next. Panel IV: the subnetwork is 4-5-6-7 and covers all individuals, so the search terminates. Note that in this case subnetwork 1-2-3-4 is equally small and covers all individuals, but would not be found by the minimum distance search starting at node 4.

The search proceeds, only in components that contain sufficiently many individuals, by recursively building up subnetworks starting from a single node (see Figure 4.4). The basis for recursion is as follows: in a component containing individuals $I = \{i_1, ..., i_{n'}\}$, then given a subnetwork *G*' of that component containing individuals $J = \{j_1, ..., j_m\}$ $(m < n')$, examine all neighbours of nodes in *G*'. Of these neighbours, find the node *g* that minimises the sum:

$$\sum_{i \in I \setminus J} d_i(g).$$

Form a new subnetwork, *G*'' by adding *g* to *G*', and repeat until all *n*' individuals are incorporated.

At each step, there may be a tie amongst neighbours for the smallest minimum distance sum, and in this case each alternative *G*'' is explored in turn. In practice this occurs frequently, leading to many calls of the recursive function in what is effectively a depth-first search strategy (Cormen 2001).

The results of this search depend on the starting node chosen, so the approach taken is to use all nodes in a component as starting nodes. This is not as costly as it may first appear owing to several steps that are taken to ensure the search runs efficiently:

- To avoid duplication of effort, a list of all subnetworks explored is kept. Suppose a search starting at node g_1 adds node g_2 first. If then the search starting at node g_2 were to add node g_1 first, the search will stop because the subnetwork g_1-g_2 has already been explored.
- The size, *s*, of the smallest subnetwork found containing all reachable individuals is maintained. Subsequently, node *g* will not be added to subnetwork *G*' to form *G*'' unless *g* is within distance $s - |G'|$ of one of the individuals needed by *G*'.
- Nodes are used as starting nodes in order (from smallest to largest) of their total minimum distance to all reachable individuals so that the smallest subnetwork size is found as quickly as possible.
- The number of calls of the recursive function can be limited (e.g. to 1,000) from each starting node, preventing excessive worst-case running times due to many ties between neighbours. Hitting 1,000 calls would make it highly unlikely that our starting node is part of a small subnetwork of interest to us.

4.2.4 Multi-Minimum Distance Search

Although the minimum distance search often works well, it does not guarantee finding the optimal subnetwork. One reason for this is that no credit is given for the fact that a neighbour might extend a subnetwork towards more than one individual. If a node is labelled with individuals 1 and 2, then its neighbour *g* has $d_1(g) + d_2(g) = 2$, overstating the actual distance. Figure 4.5 gives an example where the minimum distance search would not find the optimal subnetwork, regardless of which node is used as a starting node.

The multi-minimum distance search partially addresses this problem by preferentially seeking nodes with multiple individuals during the recursive step. The search runs recursively in the same way as the minimum distance search, with the only difference being the definition of the distances used in the sum when selecting the best neighbour of a subnetwork.

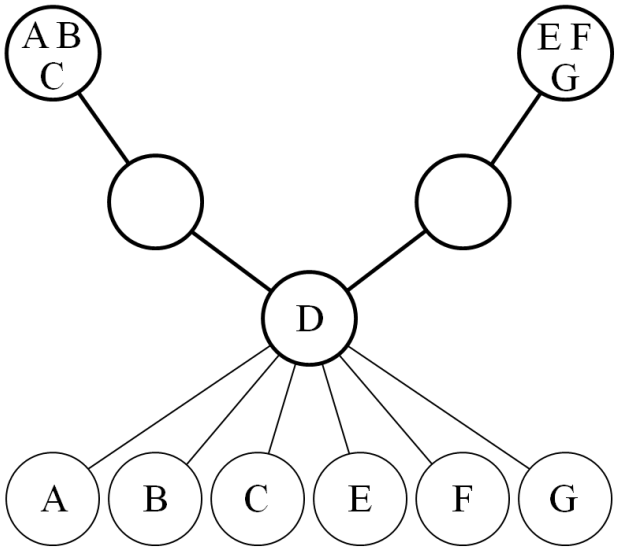


Figure 4.5 – Example of a network where minimum distance search fails
In this network the minimum distance search would fail to find the smallest subnetwork containing all individuals A–G. Nodes are labelled with the individuals attached to them. The true optimal subnetwork comprises the uppermost five nodes, indicated in by thicker lines. However, nodes from the bottom row will be incorporated into any subnetwork found by the minimum distance search, regardless of which node the search starts from. Note that in this case the multi-minimum distance search would find the optimal subnetwork.

Having previously defined the simple minimum distance $d_i(g)$, we now introduce the multi-minimum distance $d_{i,k}(g)$ for $k = 1, ..., n$. This is defined as the length of the shortest path from g to any node that is marked with $\geq k$ individuals, such that one of those individuals is individual i. If no such marked node exists (that is, the component under consideration does contain individual i, but not in any node with $\geq k$ individuals), then we set $d_{i,k}(g) = d_{i,k\text{-}1}(g)$. This is always well-defined because $d_{i,1}(g)$ is equivalent to the simple minimum distance $d_i(g)$. Distances are calculated in the same way as the simple minimum distance, for one value of k at a time, starting with $k = 1$.

The search proceeds recursively as before, only now given a subnetwork G' containing individuals $J = \{j_1, ..., j_m\}$ $(m < n')$, the next node g is the neighbour that minimises the following sum:

$$\sum_{i \in I \setminus J} d_{i,n'-m}(g).$$

The reasoning behind this approach is that when few individuals have been found, it is beneficial to extend the subnetwork towards nodes with multiple individuals. Conversely, later in the search, if only one more individual is sought, the nearest node that contains it will do.

Although this search recognises that nodes with multiple individuals are important, it is not always efficient. For example, suppose individual *i* is one of three individuals still needed by a subnetwork. The distance $d_{i,3}(g)$ does not necessarily give the distance from node *g* to a node containing those three individuals, but just the distance from *g* to a node containing *any* three individuals, one of which is individual *i*. To know the former distance would effectively require the distances be recalculated at each recursion. This is prohibitively expensive, yet provides no guarantee of finding the optimal subnetwork.

If there is a small subnetwork in which two or three nodes contain most of the individuals (which is feasible biologically), the multi-minimum distance search is likely to find it, albeit a simple extension of the minimum-distance search.

4.2.5 Program Output and User Options

Sometimes several minimal subnetworks are found that overlap (have nodes in common). Suppose subnetwork g_1-g_2-g_3 is the true underlying cause of a disease, and all individuals have a mutation in one of these genes. It could be the case, by chance, that some of the individuals also have mutations in a connected gene g_4, such that g_1-g_2-g_4 also covers all individuals. Equally, it could also be true that elsewhere in the network, three different connected genes g_5-g_6-g_7 cover all individuals by chance. BioGranat-IG can group overlapping subnetworks and return the resulting "regions" of nodes. In this case, two regions would be returned, g_1-g_2-g_3-g_4 and g_5-g_6-g_7 (along with the frequency of inclusion for each node, to quantify its importance in the group; the first group here would have a count of 2 for g_1 and g_2, and 1 for g_3 and g_4). Thus, we can provide candidate genes for experimental follow-up studies to determine the true disease-causing genes.

In BioGranat-IG, the criteria for "optimal" subnetworks can be relaxed by tolerating more nodes, or fewer individuals, up to user-specified limits (specified by *size flexibility* and *number flexibility* parameters, respectively). This flexibility allows the user a fuller analysis of potentially interesting results. In addition, there is a parameter for maximum subnetwork size, which will limit the size of any subnetworks found. This is useful in the situation where the smallest subnetwork containing all individuals is large: we might be interested in whether there exist much smaller subnetworks that contain most (rather than all) of the individuals.

In the typical situation where a small subnetwork is found that covers all individuals, BioGranat-IG offers the functionality to test whether this subnetwork is a significant finding. This can be done by generating random gene lists having the same number of genes in the network as the original gene lists. For each random instance the searches are performed, and the significance of the original subnetwork found can be

measured by the frequency with which equally small or smaller subnetworks cover all individuals in the random simulations.

It is worth mentioning at this stage that we can still construct examples of networks labelled with individuals for which none of the methods described would find the optimal subnetwork (for example, see Appendix B). But these counter-examples are much larger and more contrived, and it would seem unlikely that such a region would be biologically relevant.

4.2.6 Interaction Networks

All testing is performed in the HuPPI2 protein interaction network (PIN), as described in chapter 2. The user can choose which network to use with BioGranat-IG; as with other network-based analysis methods, this choice involves consideration of competing factors. There is typically a trade-off between network coverage (number of genes represented in the network) and the degree of confidence that can be placed on network interactions. The choice of network will also be influenced by whether a particular type of genetic mechanism is predicted and by the sequence data available. When sequencing small groups of affected individuals, one approach that could be taken is to run BioGranat-IG on smaller high-quality networks initially (this minimises the risk of connecting the individuals using false positive interactions, as would be more likely in a larger network) and proceed to larger networks if no positive results are found. Smaller networks also have the added advantage of reduced computation time.

4.2.7 Performance Testing: Methodology and Metrics

All tests examine how well BioGranat-IG can recover a specified gene complex in 1,000 tests using simulated whole exome sequencing output. For each simulation, we randomly generate lists of variant-containing genes for a fixed number of individuals. Unless otherwise stated, we use 15 individuals per simulation, to represent a typical exome sequencing study size. Each simulated individual is generated by randomly picking one gene from within the complex of interest and a fixed number of non-causal genes from the rest of the HuPPI2 network. Unless otherwise stated, we generate 35 non-causal nodes per individual. This number corresponds to the typical number of candidate genes per individual generated by exome sequencing (after filtering) (Ng et al. 2009; Simpson et al. 2011) that map to HuPPI2. We refer to this process as generating *random individuals* and *spiking in* the complex of interest.

For illustrative purposes, suppose we choose to spike in a complex of five genes. Random selection with replacement gives no guarantee that all five will be spiked into a

given set of 15 individuals (in fact, the probability is only 0.83). We do not force all five to be spiked in, as there would be no such guarantee in practice with real patients' data. Therefore, one of the metrics we look at in each test is the *number of nodes actually spiked in*.

The key result is the *number of spiked nodes recovered*, which is the number of genes from the complex of interest that are returned in the output of BioGranat-IG. This can exceed the number actually spiked in. For example, if only four of a complex of five nodes are actually spiked in, it is possible that the fifth "true" node could be found as a "jump".

In addition, we consider the *number of false nodes returned*. These are nodes that do not form part of the complex of interest, but are nevertheless returned in the output of BioGranat-IG. This can occur when, by chance, nodes neighbouring the complex are marked with individuals in such a way that an alternative smallest subnetwork can be formed by excluding one of the "true" genes in the complex, and including the non-causal neighbour. Because BioGranat-IG has no prior knowledge of the "true" disease-linked genes, all genes found are returned.

All numbers referred to in the results section are the average values found in 1,000 simulations.

4.3 Results

In this section, we firstly demonstrate that the principle underlying BioGranat-IG is sound and that the program produces valid results, using two diseases known to be genetically heterogeneous. We show that BioGranat-IG can recover the genes responsible for acne inversa (AI; OMIM #142690) and pseudohypoaldosteronism type I (PHA-I; OMIM #264350) using simulated whole exome sequencing output. We then examine the performance of BioGranat-IG under various conditions, including the nature of the underlying disease complex and the amount and quality of input data.

4.3.1 BioGranat-IG Recovers Acne Inversa Genes

AI, an inflammatory skin disease, has been shown to result from a mutation in the γ-secretase complex comprising the genes *APH1A*, *NCSTN*, *PSEN1* and *PSENEN* , with mutations in three of these genes having been directly linked to AI (Wang et al. 2010a).

Using BioGranat-IG on this complex, we were able to recover all four genes in 957 of the 1,000 simulations, but a more detailed examination of the results gives more insight into the performance.

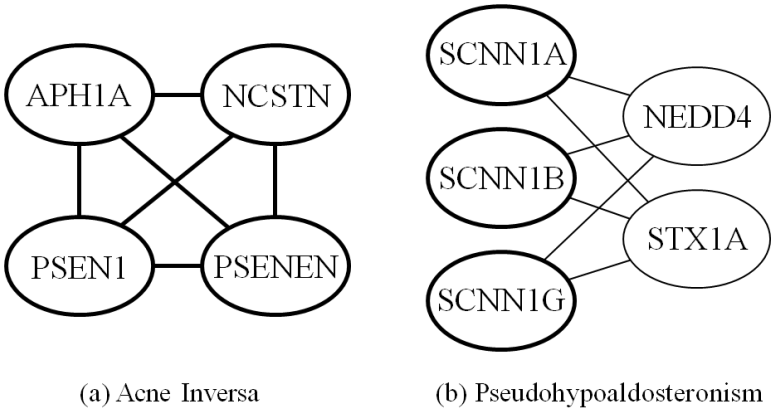


Figure 4.6 – AI and PHA-I genes in HuPPI2
The subnetworks in HuPPI2 that contain the genes responsible for the two positive control diseases. (a) AI has four underlying genes that form a clique; (b) PHA-I has three underlying genes (thick lines) that are not connected to each other but can be connected via one of two connecting genes (thin lines).

Of the 1,000 simulations, 43 had three or fewer of the four AI genes: on average the number of nodes actually spiked in was 3.957. Of these 43 simulations, none resulted in the recovery of all four AI genes. However, for every simulation, every gene that was actually spiked in was recovered, so the number of spiked nodes recovered was also 3.957. Whether any unspiked "true" nodes are recovered depends on the topology of the network around the spiked nodes. In the case of AI, the four genes form a clique in HuPPI2, and consequently, any three of the four form a connected subnetwork (see Figure 4.6a). So if only three of the genes are marked with individuals, there is no need for the fourth to be incorporated into the optimal subnetwork as a jump.

The average number of false nodes found was 0.021 (one false node in 21 of the 1,000 simulations). For AI, then, BioGranat-IG would be highly likely to direct the user towards the true causal complex, with minimal erroneous results.

4.3.2 BioGranat-IG Recovers PHA-I Genes, plus Jumps

The second disease used as a positive control is PHA-I, a disorder of electrolyte metabolism which develops in infancy. Clinical symptoms of the systemic salt loss caused by PHA-I include dehydration, respiratory problems and a high risk of life-threatening salt-losing crises (Riepe 2009). Studies have shown that the severe autosomal recessive form of PHA-I can be caused by homozygous or compound heterozygous mutations in any of the epithelial sodium channel genes *SCNN1A*, *SCNN1B* and *SCNN1G* (Riepe 2009).

In HuPPI2, these genes are not directly connected; although there is evidence for direct protein interactions (Canessa et al. 1994; Firsov et al. 1996), the databases underlying HuPPI2 do not list two publications to support this fact and hence the interactions do not

meet the quality criteria for inclusion in HuPPI2. However, all three genes are connected to *NEDD4* and to *STX1A* (see Figure 4.6b).

The average number of nodes actually spiked in was 2.990: all three genes were spiked in for 990 of the 1,000 simulations. In 989 of these, all three were correctly recovered by BioGranat-IG.

In one case, all three PHA-I genes were spiked in but only two were recovered. This occurred because 2 of the 15 individuals were spiked with gene *SCNN1G*, but for both of these individuals, *STX1A* happened to be one of the 35 randomly-generated non-causal variants. As they are not directly connected, any subnetwork including all three PHA-I genes requires at least four nodes, but in this case, a smaller subnetwork can be formed from *SCNN1A*, *SCNN1B* and *STX1A*, so this smaller subnetwork is returned. Owing to this one simulation run, the average number of spiked nodes recovered was lower than the average number of nodes actually spiked, at 2.989.

Because an additional node is always needed to connect the PHA-I genes, and there are two alternative ways to do this, both *NEDD4* and *STX1A* are always returned with the PHA-I genes found (other than the anomalous case described). Therefore, the average number of false nodes found is relatively high at 2.009 (this includes 10 simulations where an additional false gene was returned). However, this is a positive result. It shows that BioGranat-IG can work successfully even when the network used does not contain the true causal genes as a connected complex. Note that returning a small number of false genes and/or jumps is not hugely problematic, as BioGranat-IG is intended to be used to highlight genes for further experimental investigation.

4.3.3 The Effectiveness of BioGranat-IG Depends on a Number of Conditions

Having used "real" diseases to show that BioGranat-IG can find sources of genetic heterogeneity, we now examine the performance of BioGranat-IG under various conditions using artificial data.

4.3.3.1 Smaller, Less Connected Complexes Give Better Results

The ability of BioGranat-IG to find a spiked-in complex and the number of false genes it is likely to return depend on both the size of the complex and the local network topology.

To measure this, we identified three complexes of seven nodes each in HuPPI2. Complex L-7 has low connectivity (it forms a Y-shaped "branch" with a single neighbour in the rest of the network); complex A-7 has average connectivity (each node has degree 3 or 4) and complex H-7 has high connectivity (each node has degree >25). Simulations were run

Table 4.1 – Performance testing for BioGranat-IG
All numbers shown represent average of 1,000 simulations. "Complex size" = size of complex chosen to be spiked in (number of nodes). "Actually spiked" = number of nodes in the complex picked for a given simulation (for each individual, one node in the complex is picked randomly with replacement). "Recovered" = number of nodes in the complex returned in the program output. "False positives" = number of nodes outside the complex returned in the program output. "Total nodes" = total number of nodes returned in the program output (= *recovered* + *false positives*). Table continues onto following pages.

(a) Test performance on complexes of varying size in low-connectivity region (15 individuals, 35 false nodes per individual)

	Complex Size	Actually Spiked	Recovered	False Positives	Total Nodes
Complex L-2	2	2.000	2.000	0.000	2.000
Complex L-3	3	2.994	2.996	0.000	2.996
Complex L-4	4	3.953	3.981	0.000	3.981
Complex L-5	5	4.830	4.888	0.021	4.909
Complex L-6	6	5.590	5.741	0.081	5.822
Complex L-7	7	6.315	6.574	0.746	7.320

(b) Test performance on complexes of varying size in average-connectivity region (35 false nodes per individual)

	Complex Size	Actually Spiked	Recovered	False Positives	Total Nodes
15 individuals					
Complex A-2	2	2.000	2.000	0.000	2.000
Complex A-3	3	2.996	2.998	0.000	2.998
Complex A-4	4	3.948	3.961	0.015	3.976
Complex A-5	5	4.826	4.881	0.025	4.906
Complex A-6	6	5.604	5.789	0.111	5.900
Complex A-7	7	6.277	6.568	0.761	7.329
30 individuals					
Complex A-6	6	5.973	5.988	0.003	5.991
Complex A-7	7	6.911	6.964	0.017	6.981

Table 4.1 (continued)

(c) Test performance on complexes of varying size in high-connectivity region (35 false nodes per individual)

	Complex Size	Actually Spiked	Recovered	False Positives	Total Nodes
15 individuals					
Complex H-2	2	2.000	2.000	0.000	2.000
Complex H-3	3	2.991	2.993	0.043	3.036
Complex H-4	4	3.950	3.962	0.395	4.357
Complex H-5	5	4.823	4.847	1.264	6.111
Complex H-6	6	5.588	5.660	4.665	10.325
Complex H-7	7	6.322	6.297	4.657	10.954
30 individuals					
Complex H-6	6	5.977	5.984	0.401	6.385
Complex H-7	7	6.938	6.938	0.943	7.881

(d) Test performance with varying number of individuals (complex A-5, 35 false nodes per individual)

	Complex Size	Actually Spiked	Recovered	False Positives	Total Nodes
5 individuals	5	3.388	3.611	1.682	5.293
10 individuals	5	4.466	4.669	0.163	4.832
15 individuals	5	4.842	4.900	0.028	4.928
20 individuals	5	4.931	4.961	0.007	4.968
25 individuals	5	4.985	4.991	0.006	4.997
30 individuals	5	4.994	4.995	0.000	4.995

(e) Test performance with varying stringency of filtering (number of false nodes per individual) (complex H-7, 15 individuals)

	Complex Size	Actually Spiked	Recovered	False Positives	Total Nodes
15 false nodes	7	6.307	6.332	2.068	8.400
25 false nodes	7	6.332	6.325	3.359	9.684
35 false nodes	7	6.300	6.274	4.644	10.918
45 false nodes	7	6.328	6.287	6.395	12.682
55 false nodes	7	6.270	6.183	7.761	13.944

Table 4.1 (continued)

(f) Test performance when each individual is not guaranteed a mutation in the complex, but has one with a fixed probability (complex A-5, 15 individuals, 35 false nodes per individual)

	Complex Size	Actually Spiked	Recovered	False Positives	Total Nodes
No limit on subnetwork size					
Probability 0.5	5	3.977	2.241	27.565	29.806
Probability 0.6	5	4.257	3.227	22.211	25.438
Probability 0.7	5	4.474	3.953	18.241	22.194
Probability 0.8	5	4.614	4.392	13.618	18.010
Probability 0.9	5	4.742	4.718	7.352	12.070
Probability 1.0	5	4.806	4.878	0.033	4.911
Subnetworks limited to size 10					
Probability 0.5	5	3.944	1.931	13.946	15.877
Probability 0.6	5	4.251	2.951	12.890	15.841
Probability 0.7	5	4.463	3.769	10.731	14.500
Probability 0.8	5	4.622	4.421	9.248	13.669
Probability 0.9	5	4.747	4.723	5.896	10.619
Probability 1.0	5	4.832	4.895	0.054	4.949

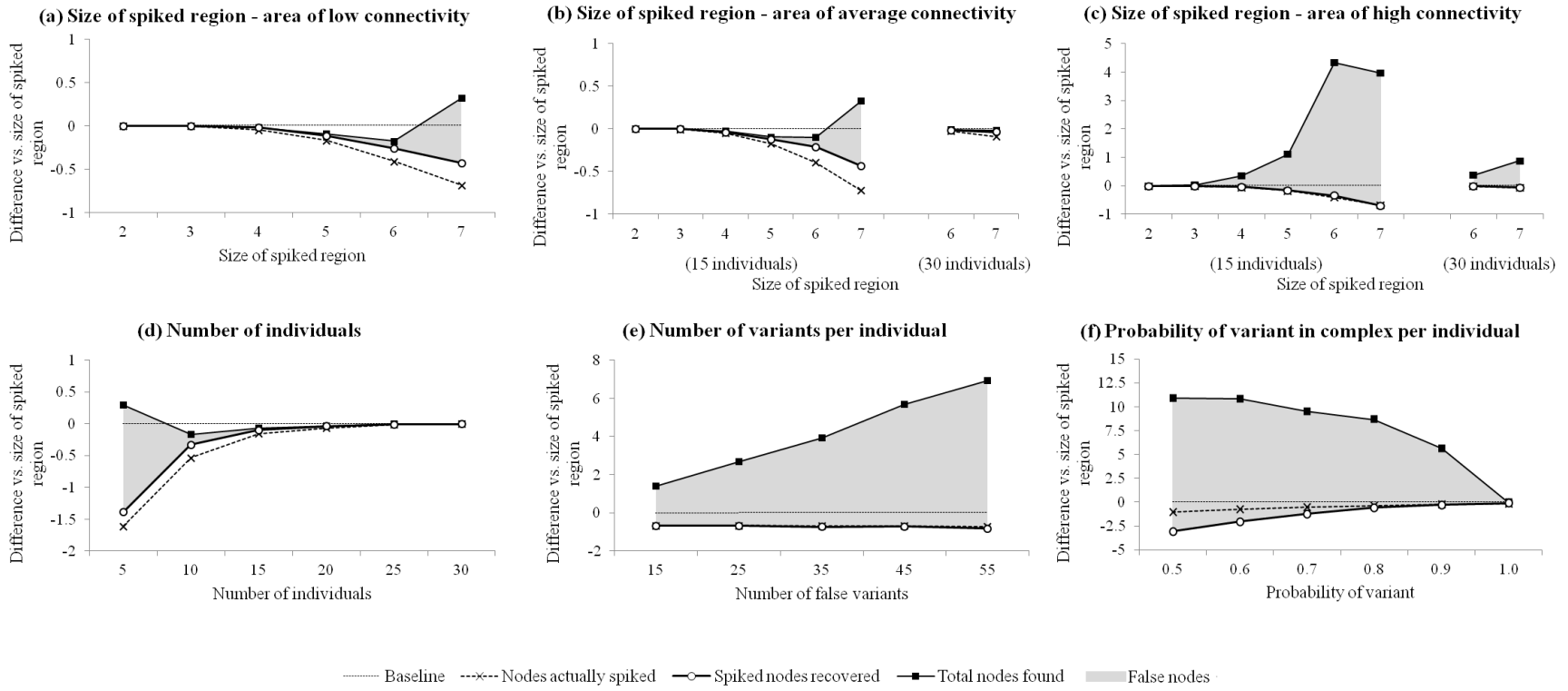


Figure 4.7 – Performance of BioGranat-IG on simulated data in various scenarios

For each graph, the vertical axis gives the value of each metric against the size of the complex being spiked in, and metrics represent the average value observed in 1,000 simulations. For example, if the spiked-in complex has four nodes and BioGranat-IG recovered 3.75 of these on average over 1,000 simulations, this would be displayed at -0.25 on the vertical axis. Each graph shows a baseline at 0. Note that graphs have different scales. (a) shows the ability to recover a spiked complex that falls in an area of low connectivity in the underlying network, for various complex sizes; 15 individuals, 35 false nodes per individual; (b) shows the same in an area of average connectivity, for 15 individuals and 35 false nodes per individual. Also shown is the improved performance for larger complexes achieved by increasing the number of individuals to 30; (c) shows the same in an area of high connectivity; (d) shows the effect of using a different number of individuals. Complex is size 5, average connectivity, with 35 false nodes per individual; (e) shows the effect of using a different filtering stringency (i.e. number of false nodes per individual). Complex is size 7, high connectivity, with 15 individuals; (f) shows the effect of changing the probability that each individual has a variant in the complex. Complex is size 5, average connectivity, 15 individuals, 35 false variants per individual, with the maximum subnetwork size limited to 10.

using subcomplexes of between two and seven nodes from each of these complexes (for example, A-5 being the subcomplex of A-7 that has five nodes).

Results for the low-connectivity complexes were excellent, with frequent recovery of additional nodes from the complex not actually spiked in, and few false nodes returned (see Table 4.1a and Figure 4.7a). For L-7, on average, only 6.315 nodes from the complex are actually spiked in, but 6.574 are recovered. In addition, only 0.746 false nodes are returned. The performance improves as the size of the complex gets smaller. This is probably owing to the reduced chance of alternative smallest subnetworks forming elsewhere by chance and reduced chance of a "true" node not actually being spiked in.

For the average-connectivity complexes, the performance suffers a little because having more nodes neighbouring the complex gives more opportunities for false variants to occur by chance in close proximity to the "true" disease nodes, thus offering alternative ways to form small subnetworks. However, the number of nodes recovered for the A complexes is broadly in line with the L complexes, while the numbers of false nodes returned are only slightly higher and still tolerable in practice (see Table 4.1b and Figure 4.7b).

The results for the high-connectivity complexes show a more substantial impact due to the presence of so many more neighbours, all of which can potentially be included as false variants (see Table 4.1c and Figure 4.7c). The power to recover genes in the H complexes is slightly reduced relative to the A complexes (for H-7, 6.322 nodes are actually spiked in but only 6.297 recovered). This reflects the increased "noise" around the complex making the true signal harder to detect. However, the bigger impact is to the number of false nodes returned, which is, for example, 4.665 for H-6 and 4.657 for H-7 (intuitively this would increase with the size of the complex – it most likely does not owing to the particular topologies of H-6 and H-7).

In summary, when the underlying complex happens to be in a highly connected part of the network, the total number of nodes found is generally higher. Because the goal is to find an unknown complex, its connectivity is not something that will be known to the experimenter *a priori*, but fortunately sequencing more individuals can help combat this problem. We tested H-6 and H-7 again, this time using 30 individuals, and the results were much improved (less than one false node found on average in each case).

4.3.3.2 Sequencing More Individuals Can Improve Results

To fully characterise the relationship between performance and number of individuals, we ran simulations with varying number of individuals on complex A-5 (to represent a typical complex). The results, given in Table 4.1d and Figure 4.7d, confirm that a

higher number of individuals leads to increased power to detect the spiked complex and fewer false nodes being returned. There are two reasons for this: the increased chance of having all nodes in the complex actually spiked in, and the reduced chance of sufficient false nodes occurring in the nodes neighbouring the region to offer alternative small subnetworks.

Clearly, the conclusion that can be drawn here is that increasing the sample size increases the power to detect a true disease-linked signal. This suggests a strategy that could be followed when BioGranat-IG is used in practice: if the number of genes returned is large, sequencing further individuals will help to narrow this list down if there is a true underlying cause of locus heterogeneity.

4.3.3.3 Stringency of Variant Filtering Affects Performance

As previously described, it is common practice when searching for genetic causes of rare diseases using whole exome sequencing data to filter out variants found in patients based on a number of criteria (e.g. genes that are well understood or in which variants are seen frequently). More stringent filtering should reduce the number of false input nodes per individual, which in turn affects the performance of BioGranat-IG.

We tested this by running simulations on complex A-5 with varying number of false nodes per individual, but for this complex, we found a marginal effect (results not shown).

The effect can be seen more clearly in the equivalent simulations run on complex H-7 (see Table 4.1e, Figure 4.7e). As expected, for 35 false variants, the results are close to what we saw for the same complex in section 4.3.3.1. With a change in the number of false input nodes per individual, there is again a marginal effect on the number of spiked nodes recovered, ranging from 6.332 with 15 false variants per individual down to 6.183 with 55 false variants. But the change in the number of false nodes returned is more dramatic: at 15 false variants per individual, only 2.068 are returned on average; this rises to 7.761 at 55 false variants per individual.

This confirms the intuitive notion that if there is a true disease-linked complex, filtering out more false nodes from the input gene lists will result in fewer false positive results in the BioGranat-IG output. Of course, the user should also be aware of the risk of erroneously filtering out variants that form part of the true complex.

4.3.3.4 There is Less Chance of Finding the "True" Complex if Some Individuals Lack a Variant

It is of course possible that the whole exome sequencing data will not contain a variant for every affected individual in what is nevertheless a true mechanism of locus heterogeneity. For example, this could be due to alternative disease pathways not present in

the network, data problems (such as incorrect base calling or incomplete exome sequencing) or some individuals being phenocopies (exhibiting a similar phenotype due to environmental effects). We simulated this on complex A-5: along with 35 false nodes, each individual received a random node from complex A-5, with probability *p*. We tested a range of *p* from 0.5 to 1.0.

Note that this also provides a good model for the situation where a true disease-linked complex is only partially represented in the underlying network. This could be the case because a true gene is not present in the network or equally because true genes are disconnected in different network components or different regions of the same component (these are problems common to many network-based analysis methods).

Initially, subnetworks of any size were allowed (see Table 4.1f). We found that as *p* decreases, it is more difficult for BioGranat-IG to pick out the true underlying complex, leading to relatively poor recovery of spiked nodes for low *p*. Worse still, the number of false nodes found grows quickly (e.g. to over 27 at $p = 0.5$).

However, better results are obtained by limiting the maximum subnetwork size (see Table 4.1f and Figure 4.7f). Without this limit, BioGranat-IG finds subnetworks containing all 15 individuals, no matter how big they may be. But it would be unreasonable in practice to expect that a large subnetwork identified using just 15 patients would be a true mechanism of genetic heterogeneity.

4.4 Discussion

We have presented BioGranat-IG, a software tool for the analysis of whole exome sequencing data with the aim of identifying groups of genes in interaction networks collectively responsible for causing a disease through locus heterogeneity.

The tool addresses the problem where several patients affected by a rare monogenic disease are exome sequenced, but no single gene is found to harbour a sequence variant for all patients. It would be possible to solve the minimal set covering problem, without using a gene network, to find the smallest number of genes across which all patients have at least one variant. However, there are two advantages to using BioGranat-IG to instead perform this search within a network. Firstly, the resulting subnetwork will be made up of genes that have already been shown to interact, so is more likely to be biologically meaningful. Secondly, the number of patients needed for results to be significant is lower in the network context, where significance is measured as the likelihood of finding an equivalently small covering set of genes by chance (see Table 4.2).

Table 4.2 – Individuals needed for significance
Fewer individuals are needed to classify a subnetwork as significantly small using BioGranat-IG in the HuPPI2 network than by using a set cover approach (no network). For each region size, the table gives the number of individuals at which finding a region of that size in which all individuals have a variant represents a significant result; 1% and 5% significance levels are shown. Results are estimated using 10,000 random simulations for each number of individuals. For the BioGranat-IG tests in HuPPI2, each individual is randomly generated with 35 genes, and the network covers 3,666 genes. Set cover tests were performed at two levels: 3,666 total genes with each individual having 35 genes chosen at random, and 20,000 total genes with each individual having 200 genes chosen at random. This last case represents the full exome, where 200 filtered variants are typically seen. The set cover tests implemented a randomised greedy search (genes were chosen to maximise the number of individuals not yet covered, with ties broken at random, and 100 iterations used per instance). As an example, for region size 4 and using BioGranat-IG in HuPPI2, 6 individuals are needed for significance at the 1% level. This means that in the set of 10,000 simulations on 6 individuals, all individuals could be covered by 4 genes or fewer less than 1% of the time, but in the set of 10,000 simulations on 5 individuals, all individuals could be covered by 4 genes or fewer more than 1% of the time.

Region size	**Individuals needed for significance**					
	Set cover @ (20,000, 200)		Set cover @ (3,666, 35)		BioGranat-IG in HuPPI2	
	5% level	1% level	5% level	1% level	5% level	1% level
1	3	4	3	3	3	3
2	6	6	5	5	3	4
3	9	9	7	8	4	5
4	12	12	10	10	5	6
5	15	16	12	13	6	7
6	19	19	15	16	7	8
7	22	23	17	18	9	10

Using simulated datasets for two diseases, we have shown that BioGranat-IG is capable of identifying the genes known to be responsible for disease phenotype. In addition, we have shown that under a range of conditions, BioGranat-IG is generally capable of picking out a relatively small subnetwork for which further experimental investigation is likely to prove insightful. Depending on the particular disease mechanisms, it is possible to use different types of networks for the analysis. For example, to identify causal genes for metabolic diseases, a metabolic network might be more informative than a PIN.

Owing to the highly interconnected nature of interaction networks, we have seen that false positive genes can be suggested by BioGranat-IG, particularly when a causal gene complex is not fully contained in the network or when there might be alternative disease pathways. However, the number of genes returned will be relatively small compared with the number of variants identified by the initial exome sequencing, and can typically be inspected manually. In addition, BioGranat-IG provides a tool to estimate the significance of

the results and configurable parameters to allow flexibility of the subnetworks returned, and the visualisation tools in BioGranat can be used to explore the results further.

If there is a true underlying complex, sequencing more individuals should reduce the number of false positive genes returned. It is important to note that many diseases cannot be linked to only a small number of genes, in which case BioGranat-IG may not be an appropriate tool to identify causative genes. In this case, sequencing more individuals should only increase the number of genes returned, rather than focusing in on a particular region. It is possible, however, that BioGranat-IG could prove useful for complex diseases in certain cases, for example, to study high-severity/early-onset cases or patients having a particular subphenotype – in which cases single pathogenic variants could contribute a majority of disease risk. (It is important to be able to distinguish these cases from the more common multifactorial cases because appropriate treatment options may differ.) Chapter 8 will examine this possibility further by describing an application of BioGranat-IG to Crohn's disease.

In its current form, BioGranat-IG represents only a first step towards solving this problem. In particular, the methods used are likely to find the smallest connected subnetwork in which all sequenced patients have a variant, but they do not guarantee it – one avenue for further work could be to improve the algorithms to minimise the possibility of missing an optimal subnetwork. In addition, further work could be done to ascertain whether a BioGranat-IG-like approach could perform better for complexes found in highly connected regions of a network, and for diseases where there is a reasonable chance of affected individuals *not* having a mutation in a true underlying complex.

The next chapter will describe HetRank, a tool which takes an alternative approach to identifying disease genes in the presence of genetic heterogeneity, and which aims to overcome some of the limitations of BioGranat-IG.

5 Development of HetRank Analysis Tool

5.1 Introduction

Intersection filtering has been shown to be a successful strategy for monogenic disease gene identification, applicable to both rare inherited and sporadic disorders and requiring no prior knowledge of the disease process or a set of candidate genes (Gilissen et al. 2011; Ku et al. 2011; Rabbani et al. 2012). While the search is exome-wide, intersection filtering is attractive because the number of genes in which several unrelated individuals carry post-filtering variants will generally be small. Conversely the effectiveness of the approach can be limited by missing data, non-exonic causal variants, and as we have seen, genetic heterogeneity (Robinson et al. 2011; Boycott et al. 2013). This thesis is concerned in particular with locus heterogeneity, whereby mutations occurring in different genes can cause the same phenotypic outcome in different patients – the genes would therefore not be revealed by simple intersection filtering (Oti and Brunner 2007; McClellan and King 2010). The BioGranat-IG tool presented in the previous chapter tries to overcome this problem by effectively widening the frame of intersection from the single gene to connected sets of genes in an interaction network.

In this chapter we present an alternative approach. HetRank is a flexible analysis tool which addresses the problem of genetic heterogeneity in exome sequencing studies by incorporating information from interaction networks into a gene prioritisation framework. Networks are ideally suited to this purpose for the same reasons that they are used by BioGranat-IG: because they group together functionally-related genes without restriction to existing curated biological pathways (Lehne and Schlitt 2012), and because it has previously been observed that genes causing the same monogenic disease are more likely to physically interact than non-disease genes (Goh et al. 2007; Feldman et al. 2008). As before it is also intuitively reasonable at a molecular level because sequence variants which affect protein conformation or the binding affinity of a protein domain can change the way or extent to which gene products interact; in theory each of the interaction partners could be vulnerable to mutation and a disease phenotype could result from the disruption to their cooperative function (Barabasi et al. 2011; Vidal et al. 2011).

HetRank has several advantages over BioGranat-IG. By incorporating network information into a gene-ranking framework, HetRank retains the ability to prioritise genes that are not included in the chosen input network. It also deals explicitly with the problem

caused by hub genes in the network; in practice hub genes can occur frequently in the optimal subnetworks found by BioGranat-IG due to their connectivity rather than true disease involvement (hub genes containing no post-filtering variants themselves are often returned in optimal subnetwork as "jumps").* The flexible ranking framework allows incorporation of diverse sources of information for variant prioritisation, and removes the need to identify appropriate thresholds for filtering of variants. Finally, HetRank can incorporate healthy control exomes to address a problem common to many exome sequencing studies: that of large and variant-tolerating genes being overrepresented among prioritised variants (Fuentes Fajardo et al. 2012; Petrovski et al. 2013); this can cause false positive findings using intersection filtering, and hence also using BioGranat-IG.

There exist several variant prioritisation tools appropriate for the study of rare monogenic diseases that integrate various sources of evidence for causality (Li et al. 2012; Sifrim et al. 2012; Carter et al. 2013; Frousios et al. 2013; Robinson et al. 2013; Sifrim et al. 2013). However, where interaction data are used it is generally either to prioritise genes based on their proximity to known disease genes in the interaction network, or to allow the user to explore genes which interact with those prioritised. To our knowledge HetRank is the first approach to incorporate interaction network information directly into the gene-ranking procedure in a hypothesis-free (that is, exome-wide) manner as a means of addressing genetic heterogeneity.†

We test our new approach using a set of test data comprising 20,000 randomised exome sequences and will show in this chapter that network information can help to rank disease-causing genes highly even under conditions of high residual unknown heterogeneity.

This work has been prepared and submitted for publication. Co-author contributions are as follows: Dr Mike Weale contributed to the method development by suggesting the mechanism by which initial gene rankings are adjusted with respect to healthy control exomes; Dr Reiner Schulz, Prof. Rebecca Oakey, Dr Michael Simpson and Dr Thomas Schlitt provided general supervision and recommendations regarding the analysis and

* Results not presented here, but note for example that in the BioGranat-IG tests depicted in Figure 4.7f in the previous chapter (using complex A-5, 15 individuals and 35 false nodes per individual), when the probability that each individual has a variant in the complex is 0.5 the gene with highest degree in HuPPI2 (*YWHAG*, degree 108) occurs in the optimal subnetwork in 19.7% of tests (and this figure rises to 31.4% when BioGranat-IG searches are not limited to subnetworks of 10 genes or fewer). In chapters 7 and 8 when BioGranat-IG is used with real whole exome sequencing studies, hub-free versions of the networks are used.

† But note that since HetRank was developed, the SPRING tool for prioritising disease-causing variants in a single exome was published (Wu et al. 2014). One of the evidence types used by SPRING is proximity to other causal genes in a protein interaction network, thereby implicitly allowing for genetic heterogeneity. If no known disease-causing genes exist, a set of seed genes can be generated based on phenotypically similar disorders.

presentation of results; Dr Simpson also oversaw and contributed data for test data generation, and advised on experimental design; Dr Schlitt also initiated and obtained funding for network method development projects (including my PhD studentship). The algorithm design (other than that credited to Dr Weale above) and implementation, test data generation, analysis, and interpretation of results presented in this chapter were performed by Nick Dand.

5.2 Methods

All data analysis is performed in the R programming language (R Development Core Team 2013), using the igraph library for network analysis (Csardi and Nepusz 2006). An implementation of HetRank in R is available online (http://sourceforge.net/p/hetrank).

5.2.1 Protein Interaction Networks and Disease Subnetworks

Results in this chapter are based on performing HetRank analysis using the PINAmin2 network described in chapter 2. Test data will make use of the OMIM disease subnetworks identified in the PINA and PINAmin2 networks in chapter 3.

5.2.2 Simulation of Whole Exome Sequencing Studies

To measure the performance of the new approach, we simulate 1,000 exome sequencing studies for rare monogenic disease. As with BioGranat-IG performance testing in chapter 4, each study is "spiked" with disease-causing variants chosen to model locus heterogeneity, and we test HetRank's ability to recover the spiked genes. However, while BioGranat-IG could be tested by generating random gene lists, HetRank requires annotated whole exome sequence data and a more sophisticated test dataset is therefore needed.

Chromosome-wise random selection (without replacement) of sequence data from 388 exomes obtained through the King's College London (KCL) rare disease programme (a subset of the exomes covered by the in-house exome database) was used to generate 200 test exomes. These were partitioned into sets of 20 exomes so that when each set is used to simulate an exome sequencing study the other 180 exomes can represent healthy controls. This was repeated 100 times to give 1,000 simulated exome sequencing studies (20,000 random exomes in total). The 388 original exomes come from unrelated individuals and include some small groups of patients sharing a rare disease phenotype, although the randomisation used should be sufficient to overcome any disease-specific bias in the test dataset.

As described in chapter 2, exome variants were annotated using ANNOVAR (Wang et al. 2010c), which provided the following criteria which can be used by HetRank to rank variants: variant effect (e.g. "nonsynonymous SNV", "stopgain SNV"), Exome Variant Server (EVS) alternative allele frequency, 1000 Genomes Project alternative allele frequency. Additional annotation gave the following variant-ranking criteria, resulting in six criteria in total: zygosity, number of observations in homozygous form in the in-house exome database, number of observations in heterozygous form in the in-house exome database. In all tests: homozygous variants were ranked ahead of heterozygous variants; "synonymous SNV" was ranked behind all other types of variant effect which were equally ranked.

To simulate exome sequencing studies, one disease-causing variant was added to each of the 20 exomes in a set. These were randomly selected from 13,413 pathogenic variants in dbSNP (Sherry et al. 2001) (build 138, downloaded 19th July 2013) that corresponded to monogenic diseases in OMIM's Morbid Map (downloaded 20th March 2013) (Amberger et al. 2009) and were annotated using the same pipeline as the KCL exome data. Since dbSNP variants did not include zygosity this was randomly set to homozygous with probability 0.9 and heterozygous with probability 0.1. Gene names for the disease-causing variants were replaced in order to model locus heterogeneity.

For this purpose OMIM disease subnetworks were randomly selected from those identified in PINA and PINAmin2 in chapter 3. PINA contains 305 unique disease subnetworks of two genes (including gene pairs drawn from larger disease subnetworks), 248 of three genes and 280 of four genes; PINAmin2 contains 150 of two genes, 108 of three genes and 111 of four genes. Gene names for the disease-causing variants were replaced with gene names from the disease subnetwork with probability 1-u (captured heterogeneity) or a gene name selected uniformly at random from the whole exome with probability u (uncaptured heterogeneity). We model uncaptured heterogeneity because in practice we might expect some proportion of disease cases to be explained by reasons other than a mutation in the disease subnetwork. To simulate balanced captured heterogeneity gene names from the disease subnetwork are randomly selected with equal probability. To simulate unbalanced captured heterogeneity gene names from the disease subnetwork are used with probabilities $p_1 = 3(1-u)/4$, $p_2 = (1-u)/4$ for two-gene disease subnetworks; $p_1 = (1-u)/2$, $p_2 = p_3 = (1-u)/4$ for three-gene disease subnetworks, and $p_1 = (1-u)/2$, $p_2 = p_3 = p_4 = (1-u)/6$ for four-gene disease subnetworks.

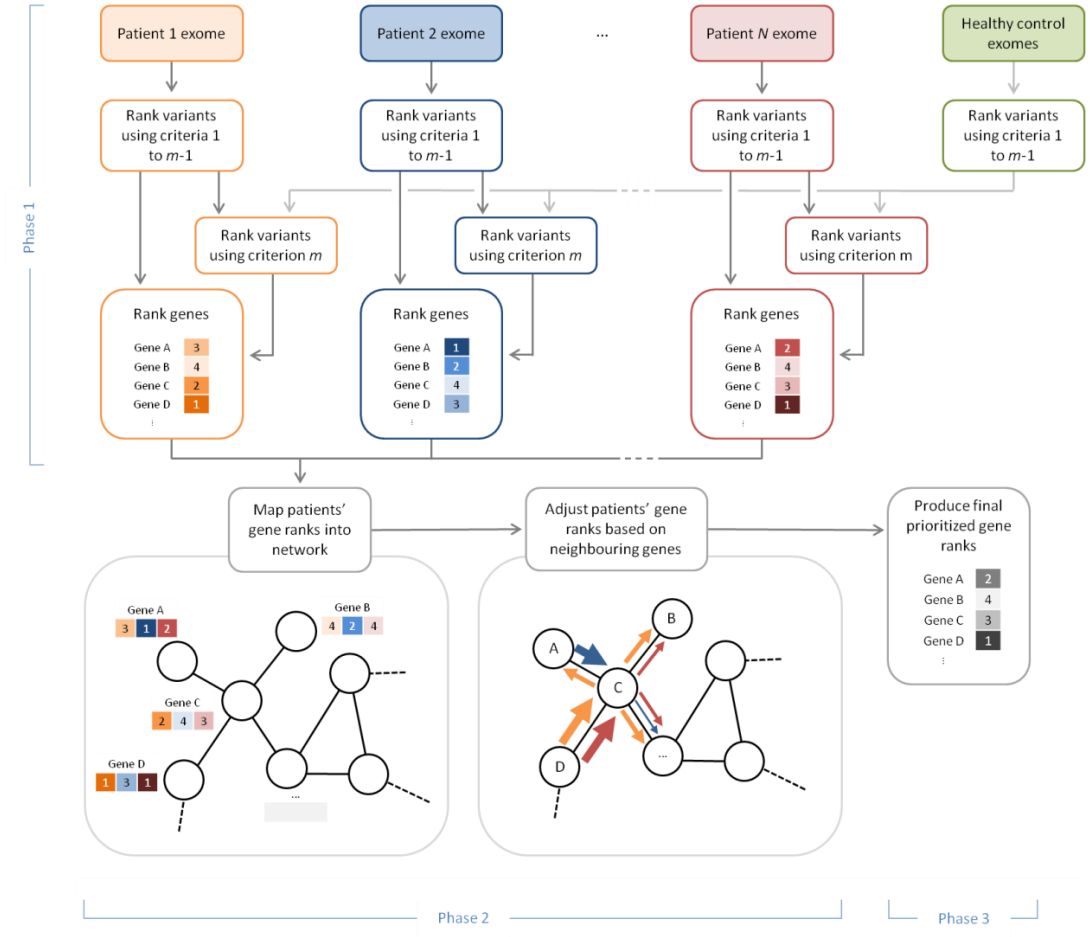


Figure 5.1 – The HetRank analysis framework
Phase 1: Variants are ranked in each affected individual's exome sequence according to a set of user-specified criteria and converted into gene ranks. Healthy control exomes can also be used to more accurately rank large and variant-tolerating genes. Phase 2: An interaction network is used to share ranking information between neighbouring genes. Neighbouring genes which rank highly in different affected individuals improve the evidence for each other's involvement in the disease process. Phase 3: Rankings are combined across all exomes in the study to give a final prioritised list of genes.

5.2.3 HetRank Gene Prioritisation Approach

Having established in chapter 3 that interacting genes cause the same monogenic diseases, we developed a method to prioritise genes for follow-up analysis in exome sequencing studies of monogenic disease, using an interaction network to overcome genetic heterogeneity (illustrated in Figure 5.1). The key concept behind the approach is to independently rank the sequence variants identified in the exomes of a number of unrelated affected individuals (according to evidence supporting each variant's disease involvement), from which a final prioritisation of genes is produced by combining rankings across the study. The final gene ranking takes into account variants found in neighbouring genes in the interaction network. This allows a gene which is not initially highly ranked in a given

individual to have its ranking improved based on evidence for disease involvement from a neighbouring gene; this acts to preferentially improve the rankings of disease genes because the sharing of evidence between any pair of interaction partners is more likely to occur consistently across unrelated individuals as a result of true locus heterogeneity than due to chance. The three key phases are as follows.

Phase 1: for each of *N* affected individuals in the study, obtain a ranking of genes according to evidence for disease involvement. First, exome sequence variants are ranked according to *m* ranking criteria with average rank being used to resolve ties. Assuming that one variant is sufficient to cause the disease in each affected individual, ranking criteria would typically include the alternative allele frequency (effectively set to zero for novel sequence variants), zygosity and variant effect (such as synonymous, missense or nonsense mutation). However, since ranking criteria are specified by the user they could also include functional prediction scores or quantification of disease-specific knowledge. Categorical criteria are ranked using values from user-supplied reference tables. Ranks are then multiplied by user-specified weights (see Discussion, section 5.4) and summed to give variant scores.

Optionally the m^{th} variant-ranking criterion can be derived from a set of healthy control exomes in order to penalise large and variant-tolerating genes that might otherwise receive an artificially high ranking. In this case the m^{th} criterion for variant *v* in gene *g* is the number of healthy control exomes in which gene *g* harbours a variant with a lower variant score (taken across the first *m*-1 ranking criteria) than variant *v*. Finally genes are ranked by the minimum variant score of any variant they contain (since a single sequence variant is sufficient to cause a monogenic disease); gene ranks are normalised by dividing by the total number of genes in which the exome carries a mutation. Genes without variants are assigned a normalised rank of 1, corresponding to no evidence for disease involvement for that affected individual.

Phase 2: network-adjusted gene rankings are obtained for each of the *N* individuals. The key assumption of our approach is that when there is locus heterogeneity, evidence for a gene's involvement in the disease process (as part of a functional pathway encoded by network interactions) can arise from a plausible sequence variant in the gene itself, or in a gene with which it interacts. For a gene *g*, the network *d*-neighbourhood, $N_d(g)$, is the set of genes which can be reached from *g* via *d* interactions or fewer (and always includes *g* itself). For individual *i* and neighbourhood parameter *d*, an adjusted ranking for each gene is calculated to reflect the fact that a better candidate disease-causing variant might exist

elsewhere in the gene's *d*-neighbourhood. The adjusted ranking for gene g is:

$$S_d(i,g) = \varphi_d(g) \times R_i(g) + \left(1 - \varphi_d(g)\right) \times \min_{h \in N_d(g)} \{R_i(h)\}$$

where $R_i(g)$ is the normalised rank of gene g from phase 1. The conservation factor $\varphi_d(g)$ is designed to limit the sharing of information based on a gene's connectedness, so that hub genes do not always receive a high adjusted ranking. It is defined as:

$$\varphi_d(g) = \frac{|N_d(g)|}{MNS(d) + 1}$$

where *MNS*(*d*) is the maximum *d*-neighbourhood size among all genes in the network. Thus a gene whose *d*-neighbourhood size is close to the biggest in the network has a high value of φ_d and its adjusted ranking stays close to its normalised rank from phase 1. On the other hand a gene with a very small *d*-neighbourhood has a low φ_d and its adjusted ranking will be close to the normalised rank from phase 1 of its best-ranked neighbour.

Since disease subnetworks are generally expected to be small (as seen in chapter 3), and for reasons of computational feasibility, the final network-adjusted gene ranking looks for better-ranked neighbours in *d*-neighbourhoods up to $d = 2$:

$$S(i,g) = \min\{S_0(i,g), S_1(i,g), S_2(i,g)\}.$$

Phase 3: network-adjusted gene rankings are combined across the N individuals in the study to give a final prioritised ranking. Although they are not p-values, the network-adjusted rankings $S(i, g)$ lie in the range (0,1] and a value closer to 0 indicates better evidence that gene g is part of a disease-causing process. Fisher's method for combining p-values $\pi_1, ..., \pi_K$ from K independent hypothesis tests has test statistic $-2\sum_{k=1}^{K} \log(\pi_k)$ (Fisher 1932). We adapt this statistic, weighting each term to reflect the fact that only some of the individuals in the study are likely to provide sequence variants that implicate a given causal gene (due to genetic heterogeneity). For each gene g, let $i_1, ..., i_N$ denote the individuals $i = 1, ..., N$ in the study, ordered such that $S(i_1, g) \geq ... \geq S(i_N, g)$ (that is, individual i_1 provides the least evidence, after network adjustment, for gene g's involvement in the disease, and i_N the most). Gene rankings are combined across individuals in the study according to $S(g) = \sum_{j=1}^{N} j log(S(i_j, g))$ (note that we do not use this value to draw conclusions about the statistical significance of our results).

Figure 5.2 illustrates how this weighted form of Fisher's sum acts when combining rankings from two exomes. Relative to Fisher's sum in its unweighted form, and weighted or unweighted simple sums, the weighted Fisher sum will prioritise genes with a very high rank

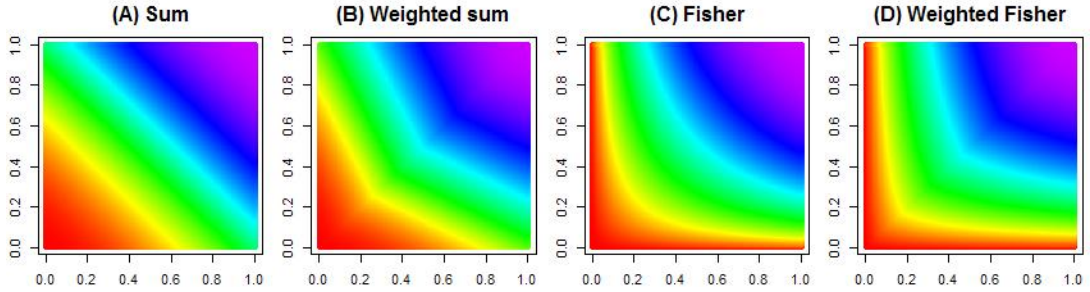


Figure 5.2 – Methods to combine rankings
Plots compare different methods to combine two normalised rankings (x-axis and y-axis values; better ranks closer to 0 at left/bottom respectively) into a final combined ranking. Two (x, y) pairs having the same colour indicate equal final rank. (A) sum: pairs (x, y) are ranked according to $x + y$; (B) weighted sum: ranking is according to $2 \times \min(x, y) + \max(x, y)$; (C) Fisher sum: ranking is according to $\log(x) + \log(y)$; (D) weighted Fisher sum: ranking is according to $2 \times \log(\min(x, y)) + \log(\max(x, y))$. Plot (D) corresponds to the approach used to combine final rankings across exomes in HetRank.

in only one of the exomes ahead of genes with moderately high rank in both. When there are more than two exomes this will mean that very high ranks in a small proportion of the exomes should be enough to rank a gene highly, which is intended to ensure that disease genes can be identified even when there is no single consistent functional mechanism underlying the disease for all cases. Results using simulated test data and the weight parameters derived in the next section confirmed that this was the most effective approach (results not shown).

Finally, all genes are ranked according to the score $S(g)$ to produce the final prioritised ranking returned by HetRank.

5.2.4 HetRank Parameters for Testing

To select appropriate weights for the tests we ran a heuristic optimisation procedure to assign weights between 1 and 8 to each ranking factor. The role of the weights is to cause the true disease-causing gene in any individual to be ranked as high as possible before the gene scores are adjusted using network information. To estimate suitable weights w_1, ..., w_7, 1,000 exome sequencing studies were simulated with a single OMIM monogenic disease gene specified (instead of sampling from an OMIM disease subnetwork), and with uncaptured heterogeneity $u = 0.5$. We sought a combination of weights that resulted in the best average rank of the disease gene across the 1,000 exome studies when HetRank was used without incorporating network information (that is, omitting phase 2).

A heuristic method was used with several steps $t = 0, 1, 2, ...$ such that at each step integer weights in the range 1, ..., $2t$ could be chosen. The optimal set of weights $w_1(t)$, ..., $w_7(t)$ at step t were "fine-tuned" by testing all $2^7 = 128$ combinations of weights

that could be derived at step t+1 by setting $w_k(t+1)$ to be either $2w_k(t)$-1 or $2w_k(t)$. Initial weights were $w_1(0) = ... = w_7(0) = 1$ and at each step t the calculation of the 7th ranking criterion (which counts how many control exomes have a smaller variant score across the first six ranking criteria in a given gene) was performed using the optimal weights from the previous step t-1. This was repeated three times, giving final weights $w_1 = w_1(3), ..., w_7 = w_7(3)$ in the range 1, ..., 8.

All ranking criteria were assigned weight 1 except for "number of observations in homozygous form in the in-house exome database" (assigned weight 8) and the 7th ranking factor which down-ranks genes ranked highly in healthy control exomes (assigned weight 8). Given their limited ability to discriminate variants the low weights for zygosity and variant effect are expected. The optimal weights also imply that the homozygous count in the in-house exome database is the most informative of the four variant frequency criteria available and strongly support the use of healthy control exomes to limit the confounding effect of large and variant-tolerant genes in the HetRank methodology.

All testing was performed using PINAmin2 as the input network for HetRank.

5.2.5 Ranking Based on Intersection Filtering

HetRank results are compared against those achieved by intersection filtering, obtained as follows. For each simulated exome sequencing study, gene lists for intersection filtering were generated for each individual by excluding synonymous variants, variants with EVS or 1000 Genomes alternative allele frequency > 0.1%, and variants in the in-house exome database (except for homozygous variants previously observed in heterozygous form). An additional gene-wise filtering step could be performed by either: excluding all genes identified by Fuentes Fajardo *et al.* as frequently enriched for false positive sequence variants in exome sequencing studies (Fuentes Fajardo et al. 2012); or: excluding genes which contain post-filtering sequence variants for ten or more of 180 healthy control exomes. Either of these additional steps represents a measure that a researcher might take to improve their intersection filtering results.

To obtain a ranking for potential disease involvement based on intersection filtering, genes were ranked according to the number of filtered gene lists in which they appear, with average rank being used to resolve ties. Gene-wise filtering using control exomes was preferred to filtering using the Fuentes Fajardo list, and to not filtering at the gene level, due to superior performance.

(This was ascertained by looking at how well intersection filtering could identify genes from three-gene disease subnetworks drawn from the PINA network, with balanced captured heterogeneity, so that $p_1 = p_2 = p_3$, and uncaptured heterogeneity $u = 0.5$. With no

gene-wise filtering, an average of 0.700 out of three disease-causing genes were ranked in the top ten across 1,000 simulated studies. This increased to 0.817 by excluding the Fuentes Fajardo genes, but using control exomes instead substantially improved this figure to 1.269.)

Genes that are excluded or contain no post-filtering variants in a study are assigned a default rank of 10,000.

5.2.6 BioGranat-IG Results for Comparison

Finally we tested whether HetRank improves upon BioGranat-IG, described in the previous chapter and to our knowledge the only previously existing tool that uses interaction networks to address genetic heterogeneity. BioGranat-IG analysis was performed using gene lists filtered as for intersection filtering, described above (except that gene-level filtering used a threshold of five, rather than ten, observations in control exomes). BioGranat-IG triplet search and heuristic (minimum/multi-minimum distance) searches were run separately using default settings (results flexibility parameters set to zero; heuristic searches limited to ten genes, 1,000 iterations per network gene and 2,000,000 iterations total). The input networks used were PINAmin2 and, considering the steps a user might take to reduce the number of highly-connected hub genes found, the hub-free version PINAmin2_d50 (constructed by removing hub genes having 50 or more interaction partners, as described in chapter 2).

5.3 Results

5.3.1 Network Information Can Improve Ranking of Disease Genes

Initially, HetRank was tested using 1,000 simulated exome sequencing studies, each "spiked" with a disease-causing variant in a gene from an OMIM disease-specific subnetwork of three genes (selected at random from those identified in Table 3.1 in chapter 3) to model locus heterogeneity. In a given study spiked with genes g_1, g_2 and g_3, each of the 20 case exomes would be assigned gene g_i with probability p_i, or a uniformly selected gene from outside the disease subnetwork with probability u that represents uncaptured heterogeneity (such that $p_1 + p_2 + p_3 + u = 1$). For each study there are 180 exomes available to act as healthy controls, with which no sequence data are shared.

Table 5.1 shows the results of testing our approach with $u = 0.5$ and the heterogeneity captured by the subnetwork split equally between the three genes ($p_1 = p_2 = p_3$). All testing was performed using the high-confidence PINAmin2 interaction network to inform gene rankings. We measure HetRank's performance by its ability to

assign high ranks to the three disease subnetwork genes. This is compared against the performance achieved using a simple intersection filtering approach, with variant- and gene-level filtering as described in section 5.2.5.

Table 5.1a shows the results obtained when disease-causing variants were assigned to OMIM disease subnetworks identified in PINAmin2, corresponding to high network coverage (interactions used to model locus heterogeneity are always included in the network we use for ranking). Comparing the final gene ranking that is achieved by HetRank after network information is incorporated against the ranking achieved using simple intersection filtering shows a consistent improvement. Of the 1,000 tests, the number in which a disease gene ranks in top position increases from 375 to 853, while the number of tests in which all three disease genes rank in positions 1-3 increases more than sixty-fold from 3 to 187. In subsequent results sections we will consider ranking in the top ten genes as a successful test since we consider ten prioritised genes to be a reasonable number for a researcher to study further in practice. The number of tests in which all three disease genes are ranked in the top ten increases from 45 to 702 when our network-informed approach is used.

Table 5.1b shows the results obtained when disease-causing variants were assigned to OMIM disease subnetworks with lower network coverage, this time identified in PINA (in this case, interactions used to model locus heterogeneity may not be covered by the network we use for ranking). Although the performance is slightly reduced, our approach still improves the number of tests in which disease genes rank in top position (from 356 to 636) or positions 1-3 (from 1 to 88) compared to intersection filtering. We now see that the number of tests in which any disease gene ranks in the top ten falls slightly, from 844 to 840. However, the number in which all three disease genes rank in the top ten still shows substantial improvement, from 35 to 349. The conclusion here is that although the power to recover the top-ranked disease gene (which should in any case be the easiest to recover without the use of network information), may be slightly reduced, our approach can clearly boost the power to recover multiple genes involved in the disease process.

5.3.2 Network Information is More Beneficial with Increased Heterogeneity

Having demonstrated that our new approach can be a valuable additional analysis tool in exome sequencing studies we examined its performance in the presence of different levels of genetic heterogeneity. In each scenario we simulated 1,000 exome sequencing studies using randomly selected OMIM disease subnetworks of a fixed size (two, three or four genes) to simulate captured heterogeneity, and a fixed level of uncaptured heterogeneity, u (20%, 40%, 60% or 80%). Further, the captured heterogeneity could be balanced (each gene in the disease subnetwork being equally likely to be disease-causing in

Table 5.1 – Ability to recover disease subnetworks comprising three genes
For all tests: uncaptured heterogeneity $u = 0.5$; captured heterogeneity split equally between the three genes ($p_1 = p_2 = p_3$). "Intersection filtering" = results obtained using ranking based on intersection filtering; "HetRank" = results obtained using HetRank approach using interaction data from PINAmin2 network; Gene 1 = highest-ranked of three disease genes in results; Gene 2 = second-highest ranked; Gene 3 = lowest ranked.

(a) Results using high-coverage disease subnetworks (identified in PINAmin2 network)

	Number of 1000 simulations achieving given ranking using high-coverage disease subnetworks					
	Intersection filtering			**HetRank**		
Gene in pathway	Gene 1	Gene 2	Gene 3	Gene 1	Gene 2	Gene 3
Ranked #1	375			853		
Ranked #1-2	66			538		
Ranked #1-3	3			187		
Ranked ≤ 10	855	369	45	987	940	702
Ranked ≤ 100	979	741	246	995	993	900
Median rank	2	39.5	303.5	1	2	6

(b) Results using low-coverage disease subnetworks (identified in PINA network)

	Number of 1000 simulations achieving given ranking using low-coverage disease subnetworks					
	Intersection filtering			**HetRank**		
Gene in pathway	Gene 1	Gene 2	Gene 3	Gene 1	Gene 2	Gene 3
Ranked #1	356			636		
Ranked #1-2	54			352		
Ranked #1-3	1			88		
Ranked ≤ 10	844	356	35	840	665	349
Ranked ≤ 100	961	700	221	913	786	542
Median rank	2	40.5	309	1	4	60

each exome) or unbalanced (one gene in the disease subnetwork being more likely to be disease-causing than the others) giving a total of 24 scenarios.

All tests were performed using the high-confidence PINAmin2 interaction network to inform gene rankings, using low-coverage disease subnetworks (those identified in the PINA network meaning that some interactions may not be covered by PINAmin2). We measure the performance of our network-informed HetRank approach by its improved ability to assign a rank of ten or less to disease subnetwork genes relative to a ranking based on simple intersection filtering.

Figure 5.3 presents the results of these tests for the 12 scenarios with balanced captured heterogeneity, and the 12 scenarios with unbalanced captured heterogeneity (which give materially similar results) are presented in Figure 5.4. Indicators of performance for all scenarios are summarised in Table 5.2.

When the captured genetic heterogeneity is ascribed to only two disease genes network-informed ranking has a modest impact compared to ranking by intersection filtering. When captured heterogeneity is balanced, and uncaptured heterogeneity is low (40% or less), we observe a slight drop in the average number of genes which rank in the top ten, and a fall in the number of tests ranking both genes in the top ten. This occurs because the limited genetic heterogeneity in these scenarios makes the disease genes easier to identify by filtering based on variant properties alone, and the network generally adds confounding information (see Figure 5.3). However, we consistently see an improvement in performance in all other scenarios. Network-informed ranking has the biggest impact for mid-range values of uncaptured heterogeneity (u = 0.4 or 0.6) when captured heterogeneity is unbalanced (with in each case >200 additional tests ranking both disease genes in the top ten; see Table 5.2). This observation leads us to conclude that our network-informed approach is increasingly beneficial over simple intersection filtering as the role of genetic heterogeneity becomes more complex, although both methods' power to prioritise disease genes suffers in extreme cases.

As we consider larger disease subnetworks, the increased levels of genetic heterogeneity more clearly demonstrate the value of incorporating network information. For example, in the four-gene case the network-informed approach considerably improves the average number of disease genes ranked in the top ten (by more than one in some scenarios), as well as the number of tests in which all four disease genes rank in the top ten (see Table 5.2). This is true at all levels of uncaptured heterogeneity and whether captured heterogeneity is balanced or unbalanced. Of particular note is the case of 20% uncaptured heterogeneity. Here, network information causes an average increase in the number of disease genes ranked in the top ten of 0.93 when the captured heterogeneity is balanced, but 1.31 when captured heterogeneity is unbalanced. This difference reflects the fact that an unbalanced split of locus heterogeneity makes it harder to correctly identify the disease genes based on variant properties alone, whereas network-informed ranking is relatively robust to the allocation of disease involvement between connected genes.

For three- and four-gene disease subnetworks with 80% uncaptured heterogeneity, none of the exome studies ranked all disease genes in the top ten based on intersection filtering. The relatively modest results for these scenarios in Table 5.2 reflect the difficulty of correctly prioritising disease genes when there is a high level of genetic heterogeneity.

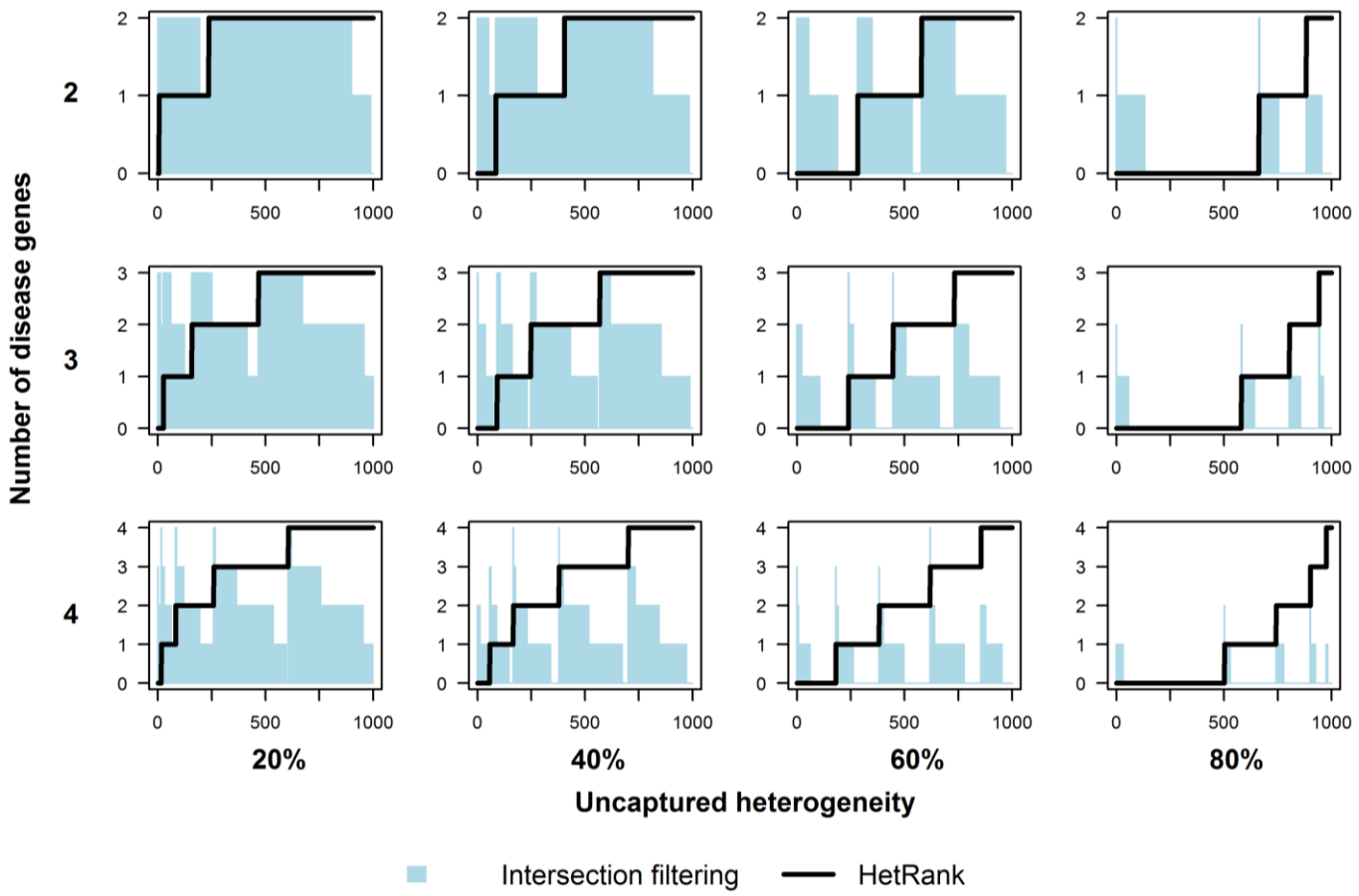


Figure 5.3 – Performance of HetRank at varying levels of genetic heterogeneity when network-captured heterogeneity is balanced
For each scenario the results of 1,000 simulated exome sequencing studies are displayed, ordered by the number of disease genes ranked in the top ten using the HetRank (network-informed) approach (black line). Blue shading indicates the number of disease genes ranked in the top ten based on a simple intersection filtering approach. Interpretation: blue shading above the black line equates to drop in performance using HetRank; white space beneath the black line equates to improvement in performance using HetRank. The net differences between these areas are presented in Table 5.2.

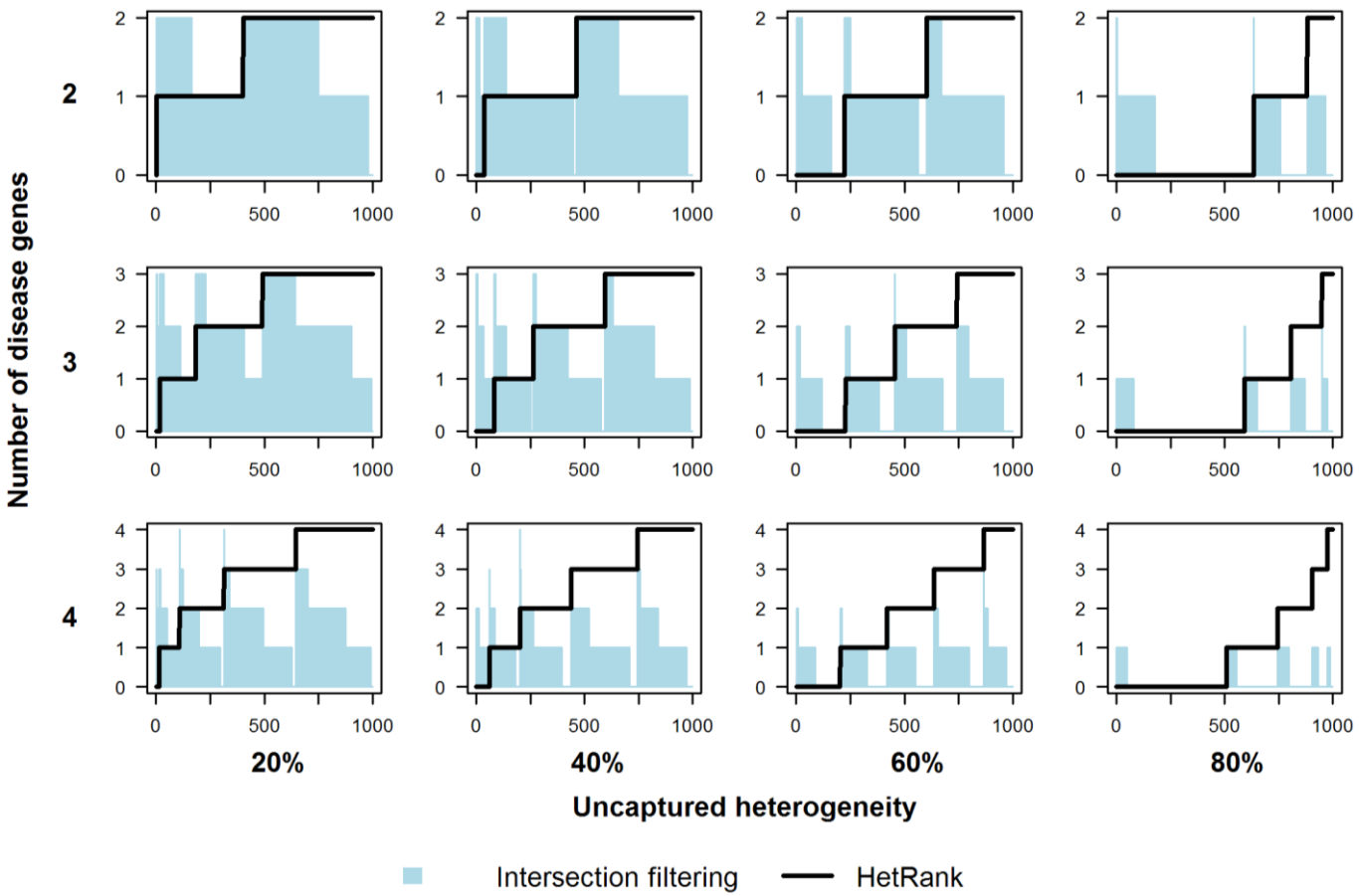


Figure 5.4 – Performance of HetRank at varying levels of genetic heterogeneity when network-captured heterogeneity is unbalanced
Plots are as in Figure 5.3 but now describe results for tests in which captured heterogeneity is unbalanced.

Table 5.2 – Improved ability to recover disease subnetworks under varying levels of genetic heterogeneity using network-informed HetRank approach relative to simple intersection filtering
All results are based on 1,000 tests.

	Average increase: # disease genes ranked in top ten				**Increase: # tests in which all disease genes rank in top ten**			
	Uncaptured heterogeneity				Uncaptured heterogeneity			
	20%	40%	60%	80%	20%	40%	60%	80%
Balanced captured heterogeneity:								
2-gene disease subnetworks	-0.08	-0.11	0.03	0.15	-92	-55	146	105
3-gene disease subnetworks	0.15	0.49	0.75	0.48	189	341	262	59
4-gene disease subnetworks	0.93	1.36	1.36	0.76	367	297	144	25
Unbalanced captured heterogeneity:								
2-gene disease subnetworks	0.11	0.24	0.20	0.09	86	227	275	113
3-gene disease subnetworks	0.36	0.57	0.73	0.42	289	344	257	53
4-gene disease subnetworks	1.31	1.39	1.20	0.69	349	255	136	27

Figure 5.5 gives an alternative presentation of the summary results, which also shows the performance that is achieved by HetRank without including the network-adjustment step (i.e. skipping phase 2 of the process; see section 5.2.3) and the performance of simple intersection filtering without performing the gene-level filtering against control exomes (described in section 5.2.5)

Again, a number of inferences can be made from these graphs, which show the average number of disease genes identified in 1,000 simulated exome sequencing studies for varying levels of genetic heterogeneity. Every line slopes downwards, due to the increased difficulty of identifying disease genes at higher levels of uncaptured heterogeneity. As expected, intersection filtering is more effective at all levels of genetic heterogeneity when gene-level filtering is performed using control exomes (dark blue line) than when no gene-level filtering is performed (light blue line). Using HetRank without performing the network-based rank adjustment (orange line) further improves performance at all levels of genetic heterogeneity, suggesting that the HetRank framework makes more effective use of sequencing data than intersection filtering. Lastly, the final network-adjusted HetRank results (green line) show that network information is more beneficial as heterogeneity increases. When captured heterogeneity is modelled by a balanced split between two genes (top-left graph in Figure 5.5), the performance is actually less good than using HetRank without the network adjustment – suggesting that network information is more confounding than informative in this scenario. However, the performance gap decreases as the level of uncaptured heterogeneity increases and, further, when the captured heterogeneity is

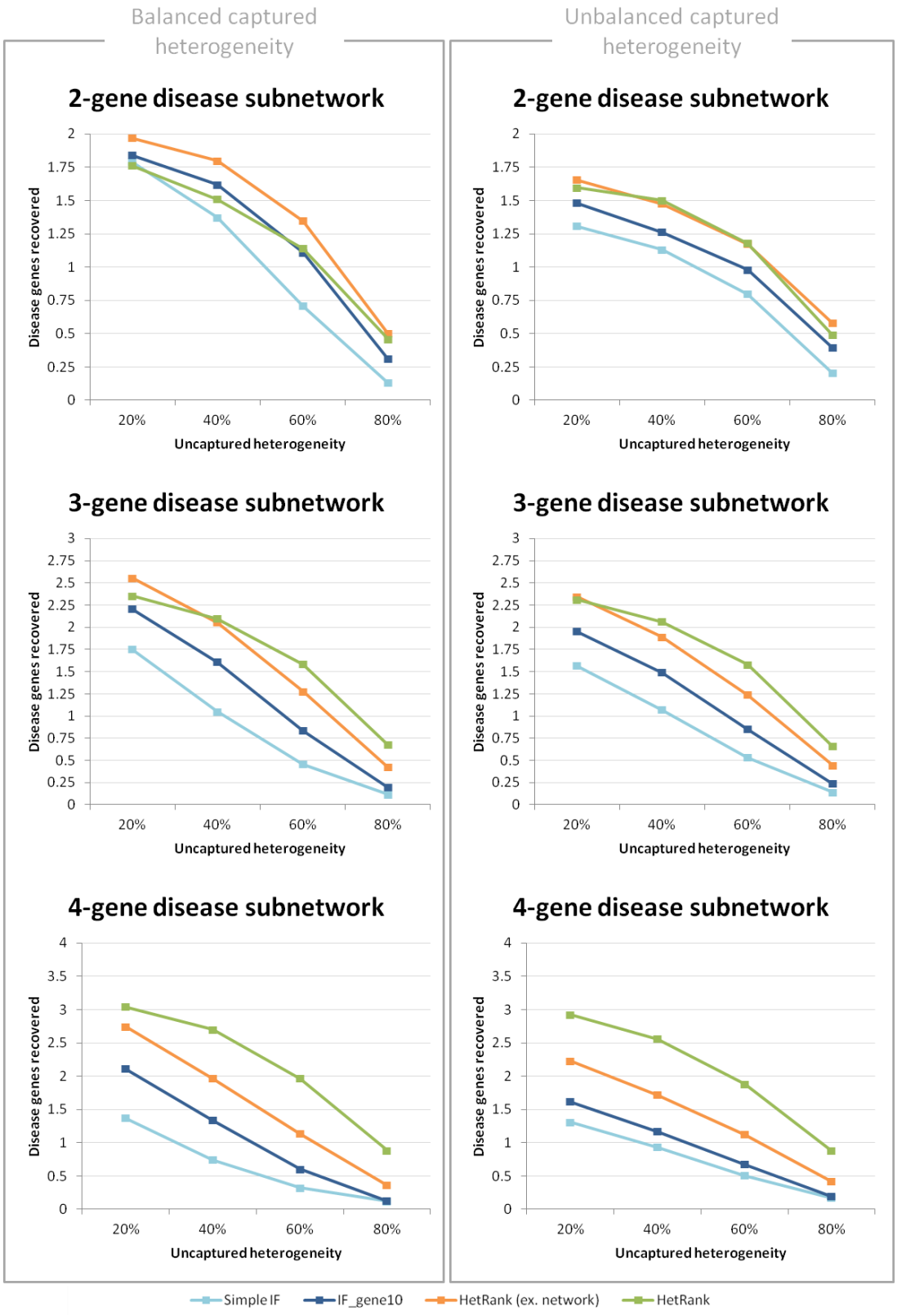


Figure 5.5 – Summary performance of HetRank at varying levels of genetic heterogeneity
See next page for full figure legend.

Figure 5.5 – Summary performance of HetRank at varying levels of genetic heterogeneity (previous page) Graphs show average number of disease genes ranked in the top ten by each method across 1,000 simulated exome sequencing studies. "Simple IF" = intersection filtering based on variant properties alone; "IF_gene10" = intersection filtering with gene-level filtering against control exomes as described in section 5.2.5; "HetRank (ex. network)" = results from HetRank if network-adjustment is not included (i.e. skipping phase 2 as described in section 5.2.3); "HetRank" = final HetRank results. "IF_gene10" and "HetRank" refer to the tests depicted in Figure 5.3 and Figure 5.4.

unbalanced (top-right graph) HetRank's performance with and without network information is similar. When captured heterogeneity is modelled by three-gene disease subnetworks (middle row of graphs), network information is beneficial at higher levels of uncaptured heterogeneity, while for four-gene disease subnetworks it improves performance considerably in all scenarios (bottom row of graphs).

5.3.3 HetRank Improves on BioGranat-IG Results

BioGranat-IG requires a list of genes that harbour sequence variants (after filtering for variant frequency and effect) for each affected individual in the study, and an input network. Each post-filtering gene is treated as equally likely to be causal and the tool outputs one or more candidate disease subnetworks. We expected the HetRank approach to outperform BioGranat-IG due to the fact that it ranks all variants for evidence of disease involvement (thus not requiring a fixed threshold) and because it incorporates network information into a prioritisation (thus meaning that even genes not featuring in the interaction network can be prioritised). HetRank also has the advantage of directly addressing problems caused by large and variant-tolerant genes, and genes which are highly-connected in the network (hubs).

Figure 5.6 compares the results of the two approaches. Since BioGranat-IG will return all the optimal subnetworks that it finds the user cannot specify the number of genes that will be returned. Therefore performance of the HetRank approach was assessed by considering the number of disease-causing genes to which it assigns a rank better than or equal to the number of genes returned by BioGranat-IG for each simulated exome-sequencing study.

Figure 5.6a shows that, for exome studies simulated to have 50% uncaptured heterogeneity and (balanced) captured heterogeneity modelled by a three-gene disease subnetwork, HetRank performed at least as well as BioGranat-IG's heuristic network search in 96.5% of simulations, and better in 72.1% of simulations. BioGranat-IG's exact triplet search failed to produce results for comparison due to the complexity of the PINAmin2 network.

Considering the steps a BioGranat-IG user might take to improve their results, we re-ran BioGranat-IG using a hub-free version of the network, PINAmin2_d50. To address

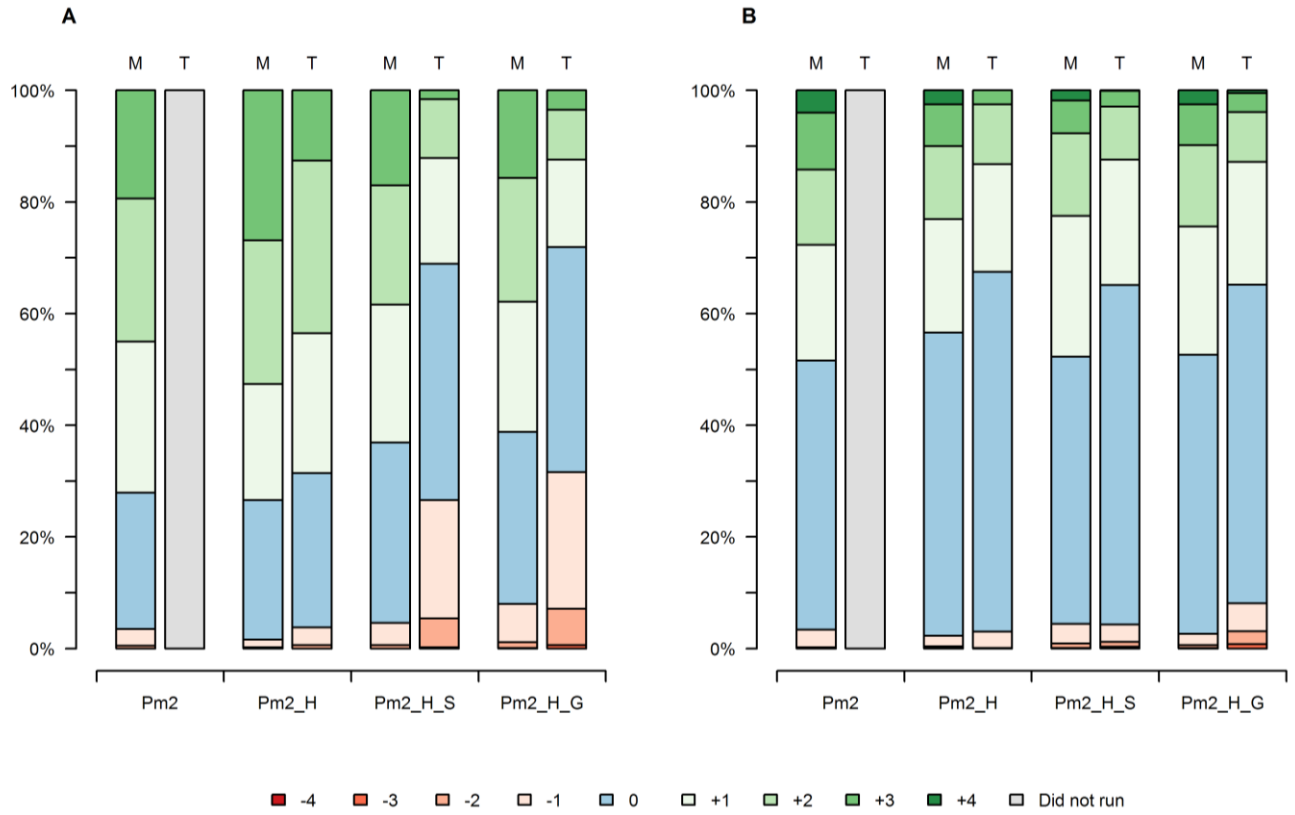


Figure 5.6 – HetRank performance compared against BioGranat-IG
Increase in number of disease genes identified by using HetRank instead of BioGranat-IG (considering same number of result genes) for 1,000 simulated exome studies. (A) three-gene disease subnetwork, balanced captured heterogeneity, 50% uncaptured heterogeneity; (B) four-gene disease subnetwork, balanced captured heterogeneity, 80% uncaptured heterogeneity. Columns represent different conditions for BioGranat-IG runs: M = heuristic (min/multi-min distance) search; T = triplet search; Pm2 = using PINAmin2 network; Pm2_H = using hub-free PINAmin2_d50 network; Pm2_H_S = using PINAmin2_d50 with additional gene list filtering against a single list; Pm2_H_G = using PINAmin2_d50 with additional gene list filtering against genes overrepresented in control exomes. Interpretation: green shading indicates HetRank improves on BioGranat-IG performance; blue shading indicates equivalent performance; red shading indicates worse performance.

the overrepresentation of highly-mutated genes we also tried using input gene lists which excluded either a fixed set of dubious genes (Fuentes Fajardo et al. 2012) or a set of genes overrepresented after variant-filtering in each study's corresponding set of healthy control exomes (as with intersection filtering, defined as containing a post-filtering variant in ten or more of 180 controls). We found that even with these modifications to the BioGranat-IG methodology, its heuristic network search was outperformed by HetRank. However, BioGranat-IG's exact triplet search (which limits itself to finding small subnetworks of interest) performed reasonably well following both network hub removal and additional gene filtering against healthy control exomes. Even in this case, HetRank performed as well or better in 68.4% of simulations.

An increased level of genetic heterogeneity results in a more marked improvement in results using HetRank compared to BioGranat-IG. Figure 5.6b shows the results of simulations where disease-causing genes were selected using a four-gene disease subnetwork and uncaptured heterogeneity of 80%. In this case, regardless of the search method used in BioGranat-IG or the further steps taken to improve its results, HetRank found as many or more disease genes in at least 91.9% of simulations. As we will discuss below it is likely that this scenario represents a truer picture for many monogenic diseases than the lower-heterogeneity scenario. With this in mind the results here show that our new approach is better equipped to deal with this genetic heterogeneity than currently available methods.

5.4 Discussion

Locus heterogeneity reduces the power of exome-sequencing studies to identify the molecular basis of a monogenic disease because it limits the expected overlap of genes harbouring deleterious mutations in unrelated affected individuals. This heterogeneity presents a challenging problem but we have shown that we can improve the prioritisation of disease genes by incorporating information from an interaction network in a hypothesis-free manner; that is, without specifying a set of candidate or "seed" genes.

Such an approach is particularly valuable when there is a high level of heterogeneity. There are currently many known sets of two to three interacting and disease-causing genes (see chapter 3), but we expect to see larger connected sets in future as new disease-causing genes are identified and as more comprehensive interaction networks are developed (Yu et al. 2011). Gilissen *et al.* suggest that there is a scale of genetic heterogeneity broadly corresponding to disease prevalence (Gilissen et al. 2011), and our results support the use of HetRank in sequencing studies at the higher end of this scale. This

could include studies of groups of patients with the same or very similar clinical phenotypes having unknown and potentially diverse molecular causes. A recent study of autosomal recessive hereditary spastic paraplegias, for example, proposed eighteen novel candidate genes where a single variant in each was thought to be disease-causing in different families (Novarino et al. 2014).

Our approach performs well in the presence of heterogeneity not captured by an interaction network (which we tested using the parameter *u*). This might include missing interaction data, disease variants that are not protein-coding (intronic or intergenic variants), as well as potential non-genetic disease causes such as epigenetic or environmental causes. As might be expected, though, Table 5.2 shows that better network coverage of interactions between disease-causing genes improves HetRank results. Our approach does not specifically require a protein interaction network be used and as such it may be beneficial to seek increased coverage of the interactome by using networks which integrate different types of gene relationships (Lee et al. 2011; Vidal et al. 2011; Khurana et al. 2013).

Careful inspection of Figure 5.3 and Figure 5.4 shows that network information does not improve the performance of all studies equally. Even at higher levels of genetic heterogeneity there are examples where more disease genes are ranked in the top ten using the simple intersection filtering method. This demonstrates what we already know: it is important that a researcher also considers the evidence for a gene's involvement in a disease independently of the genes with which it interacts. One way to do this could be to consider HetRank gene prioritisations alongside those obtained by intersection filtering when analysing exome sequencing results.

On a similar note, users of HetRank should understand the limitations of the tool. At higher levels of genetic heterogeneity in particular, we saw many simulated studies in which no disease genes could be ranked in the top ten. Furthermore, for any given monogenic disease, the model underlying HetRank (that locus heterogeneity can be at least partially explained by interacting genes) may be inappropriate; the tool itself cannot determine if this is the case. Even for an exome sequencing study in which disease-causing genes are ranked in the top ten by HetRank (our measure of success in performance tests), those ten will also include non-causal genes. A thorough examination of the top-ranked genes is likely to be needed to discern the genes of interest. However, what HetRank can do is provide the user with a starting point; by studying the high-ranking genes and the variants that they and their interaction partners contain, and combining this with existing functional annotation or disease-specific knowledge, hypotheses can emerge concerning putative disease mechanisms which can be taken forward for further testing.

An important point to note is that the use of the HetRank framework does not preclude the use of other tools designed for gene or variant prioritisation, but can be considered complementary to existing approaches. Variant effect prediction tools such as SIFT (Kumar et al. 2009) and PolyPhen (Adzhubei et al. 2010) can be incorporated into HetRank by including their prediction scores as variant-ranking criteria, and nothing prevents the same approach being taken for existing tools such as Endeavour (Aerts et al. 2006), Exomiser (Robinson et al. 2013) or CADD (Kircher et al. 2014), which themselves integrate diverse sources of evidence for deleteriousness or disease involvement. Even if a user relies entirely on an existing tool to integrate evidence sources, HetRank can still make a valuable contribution to addressing genetic heterogeneity by adjusting this evidence with reference to an interaction network.

An interesting future exercise might be to compare HetRank performance against that of SPRING, a recently developed tool which prioritises non-synonymous variants according to the likelihood that they cause a specified disease (Wu et al. 2014). Alongside direct variant effect prediction scores, SPRING uses indirect evidence for disease-causality (implicitly recognising the potential for locus heterogeneity) by looking for "association" with seed genes that cause the same or similar diseases; such association includes similarity of Gene Ontology annotation, protein sequence, protein domain annotation and curated pathway annotations, as well as proximity in a protein interaction network (Wu et al. 2014). SPRING differs from HetRank by taking a different algorithmic approach. Most notably from a user's perspective, HetRank provides complete flexibility to use customised ranking criteria and accrues evidence from the exomes of multiple affected individuals to prioritise relevant genes. Unfortunately the simulated exome sequencing studies generated in section 5.2.2 would not be suitable to assess the performance of SPRING because the real disease-causing variants on which each study's spiked variants are based are not specific to any single phenotype.

In the HetRank framework, user-specified weights are used to integrate evidence sources, and there are several considerations to bear in mind when choosing appropriate weights. The weights are applied linearly to combine rankings and can therefore be interpreted as quantifying the user's beliefs about the relative informativeness of each evidence source. That is, the ranking criteria reflect aspects of the genetic model assumed to underlie the disease and the weights reflect the user's confidence in the validity of each aspect of this model and how well these aspects are represented by the variant annotation.

In all of our simulated exome sequencing studies we used a previously determined set of weights: we chose to use integer weights between 1 and 8, and optimised these using simulations with a single disease gene (see section 5.2.4). It will usually be beneficial to give

a high weight to a measure of variant frequency since mutations that cause rare monogenic diseases are expected to be rare and in many cases private (found in a single family), and because frequency is a highly discriminative property of sequence variants. A more moderate weight is appropriate for ranking criteria that are less specific in their ability to distinguish pathogenic mutations from benign variation, such as categorical criteria with few distinct categories (like zygosity or variant effect). Care should also be taken when choosing weights for criteria that are likely to be correlated, such as different measures of variant frequency.

It is also highly recommended to give a high weight to the additional ranking factor that is automatically generated based on healthy control exomes (depending on the number of control exomes used). This ranking factor is important because it is designed to de-prioritise genes which might otherwise rank highly because they have a propensity to contain, and tolerate, sequence variants with characteristics typical of disease-causing mutations (such as low frequency and non-synonymous effect on the gene's protein product).

This chapter has presented HetRank, a flexible analysis method which uses variant-ranking in unrelated exomes to combine several sources of evidence for involvement in a monogenic disease. In an interaction network, neighbouring genes which rank highly in different affected individuals improve the evidence for each other's involvement in the disease process. The final prioritisation is obtained by combining adjusted gene ranks across all exomes in the study, and we have demonstrated using simulated data that this can effectively deal with a considerable degree of genetic heterogeneity.

An application of HetRank to a real whole exome sequencing study will form one of the analyses presented in chapter 7.

6 Supporting Methods for Application to Real Disease Data

This chapter will describe two supporting methods that will be used in the subsequent chapters, where the tools developed in chapters 4 and 5 are applied to real exome sequencing studies. Section 6.1 describes a method to prioritise BioGranat-IG optimal subnetworks for further study. Section 6.2 presents an update to an existing method, Region Growing Analysis (RGA; Lehne 2011), which takes a ranked gene list and identifies regions in an interaction network that are significantly enriched for highly-ranked genes. RGA will be used in chapter 7 for the interpretation of HetRank results, and in chapter 8 as a direct analysis tool.

6.1 Prioritisation of BioGranat-IG results

As described in chapter 4, the BioGranat-IG tool includes a permutation test that can be used to estimate the probability that for an observed optimal subnetwork, a subnetwork harbouring post-filtering variants in more exomes or a subnetwork of equal or smaller size harbouring post-filtering variants in the same number of exomes could be observed by chance. This permutation test makes the simplifying assumption that genes in the network are equally likely to contain a post-filtering variant by chance, which is untrue in practice (Fuentes Fajardo et al. 2012; Petrovski et al. 2013).

Further, BioGranat-IG can return several equivalent "optimal" subnetworks. For example, a triplet search might return several different triplets harbouring variants in five exomes, but none with variants in more than five. BioGranat-IG's significance test cannot discriminate between these results.

If any of these triplets overlap (have genes in common) then it may be instructive to merge them and consider the bigger subnetwork they generate. If not, they must be considered separately. The quadruplet search and heuristic searches will likely generate different subnetworks too. In addition, BioGranat-IG will be employed using different networks and different filtering levels, potentially generating a wide range of results. To address this, an alternative method to provide an independent measure of a subnetwork's viability (as a pathway causing a rare monogenic disease through a mechanism of locus

heterogeneity) was developed as one means of prioritising subnetworks for efficient follow-up.

The method assumes that if all variants v can be assigned an independent score $f(v)$ quantifying their likelihood of being involved in the disease, then the most promising BioGranat-IG subnetworks should be those which are enriched for high-ranking variants. In theory, the score could apply at the variant level (such as a variant effect prediction score) or at the gene level to the gene in which v is located (such as the Residual Variation Intolerance Score [RVIS][*]). The score could be specific to the disease in question (such as a measure of gene expression in relevant tissues) or could reflect general evidence for disease involvement (such as variant effect prediction scores or RVIS).

In addition to the score f, a summary statistic $F(f(v_1), ..., f(v_n))$ is required to appropriately combine the scores across variants $v_1, ..., v_n$ that are observed in the same subnetwork. For monogenic diseases where a single variant is expected to be causal for each individual, if an exome has two or more variants in the subnetwork F might incorporate only the highest scoring of those variants.

Enrichment of a subnetwork for high-ranking variants is estimated using a permutation-based approach. Given a subnetwork that contains a variant for a subset I of the case exomes, the summary statistic is calculated as follows. For each individual i in I, having $n_i \geq 1$ variants in the subnetwork, the set V_i of all variants which map anywhere in the network is identified. In each permutation, n_i variants are randomly selected from V_i for all i in I (without reference to the network), and the summary statistic is calculated. 100,000 permutations are performed and the significance of the observed subnetwork's test statistic is calculated as the proportion of permutations in which a greater or equal summary statistic is generated. This test therefore examines how enriched the observed subnetwork is for high-ranking genes in comparison to unconnected combinations of variants in the same individuals.[†]

In the following chapters the variant scores f were obtained from the variant effect prediction tool KGGSeq. KGGSeq implements a logistic regression which combines scores from SIFT, PolyPhen-2, MutationTaster, PhyloP and a likelihood ratio test based on sequence conservation (Li et al. 2012). This results in an estimate for each variant of the

[*] RVIS is a measure that estimates each gene's tolerance of functional variation based on the observed functional and non-functional variation in a sample of ~6,500 whole exome sequences (Petrovski et al. 2013).

[†] Note the test assumes that each individual has a reasonable number of variants remaining after filtering. If filtering is so effective that it leaves very few variants per individual then this prioritisation test would not be required (because BioGranat-IG would be unlikely to find many subnetworks by chance).

probability that it causes a monogenic disease. In each study KGGSeq was applied to each case exome. For variants which received no score, an estimated score based on variant consequence (e.g. "nonsynonymous SNV", "stopgain SNV", "frameshift deletion") was used, where this was calculated as the mean value of all (scored) variants across all exomes (after filtering on variant frequency only).

The summary statistic used to combine these probabilities across a subnetwork *G'* containing variants v_1, ..., v_n approximates the probability that at least one of the variants in the subnetwork causes a monogenic disease (assuming that all variants have independent probabilities of being disease-causing). It is calculated as:

$$F\big(f(v_1), \dots, f(v_n)\big) = 1 - \prod_{i=1}^{N'} (1 - f(v_i^*)),$$

where the product is taken over all *N'* individuals having a variant in a gene in *G'* and v_i^* is the variant with highest probability of causing a monogenic disease among the n_i variants carried by individual *i*. (Note that $\sum_{i=1}^{N'} n_i = n$.)

For convenience, this prioritisation method as applied to BioGranat-IG results will for the remainder of this thesis be referred to as *KGGSeq-prioritisation*.

6.2 Region Growing Analysis

RGA was originally designed to be used with genome-wide association study (GWAS) data, with the aim of identifying regions in an interaction network that are enriched for association signal (Lehne 2011). However, since the input data required are simply an interaction network and a ranked list of genes (originally the ranked list was derived using GWAS p-values), it can be appropriate to apply RGA in other contexts; for example in chapter 7 we make use of RGA to help interpret the prioritised list of genes produced by HetRank.

RGA is a bundle in BioGranat (see chapter 2). The original version of RGA was implemented by Dr Benjamin Lehne, Nikolaos Barkas and Christopher Tebbe. As part of the present work a new algorithm for RGA was devised and implemented, allowing analyses to be performed considerably faster.

Given a ranked gene list, RGA requires two threshold parameters, α and β ($\geq \alpha$). Any gene in the network whose rank is below (i.e. better than) the threshold α is considered a *seed node*. In the original implementation, network regions are "grown" around each seed node by incorporating neighbouring genes that have rank $<\beta$ (termed *member nodes*).

Optionally, "jumps" over one node with rank >β may be permitted, with jumped nodes also being added to the region. After each iteration any regions which overlap are merged (see Figure 6.1). This process is repeated for a fixed number of iterations. In practice using GWAS data, this analysis generally found one large region and a number of much smaller regions (Lehne 2011). To test whether a region is larger than expected by chance (contains significantly many seed or member nodes), network permutations can be used to generate random regions allowing an empirical p-value to be estimated. Permutations use a degree-constrained label-shuffling approach. Typically this will be repeated for a range of α and β thresholds to explore how the regions found vary.

It is very difficult to estimate accurately the time-complexity of the original RGA algorithm due to the number of conditions that can vary between scenarios, but crudely (when jumps are allowed) the running time grows as:

$$O\left(n_{perm}\left(n_{\alpha\beta}(n_V \log n_V + n_E) + \sum_{\alpha,\beta} n_{iter}\left(n_V^{\alpha,\beta} + \sum\nolimits_{v \in N_1(V_{\alpha,\beta})} d(v)\right)\right)\right). \qquad (*)$$

Here, n_{perm} is the number of network permutations used to estimate significance. $n_{\alpha\beta}$ is the number of combinations of α and β tested, and n_V and n_E are the number of nodes and edges in the network respectively. The term $n_{\alpha\beta}(n_V \log n_V + n_E)$ is included because a permuted network is generated for each (α, β) pair, requiring network nodes to be sorted by degree (with complexity $O(n_V \log n_V)$ (Oracle 2011)) and all edges to be processed; the term also accounts for locating seeds in the network, with cost $O(n_V)$. To this is added a sum across all (α, β) pairs. The terms inside the sum are due to the fact that neighbours of each node in a region, plus nodes within one jump of a region, are explored at each iteration; n_{iter} is the number of iterations specified by the user, $V_{\alpha,\beta}$ is the set of all nodes in any region for thresholds α and β, and $n_V^{\alpha,\beta}$ is the number of such nodes; $N_1(W)$ is the 1-neighbourhood of a set of nodes W (all nodes in W plus direct neighbours), and $d(v)$ is the degree of node v (it is assumed that most nodes in the region are found within the first few iterations so that most iterations search over fully- or near-fully grown regions). In this case merging of regions should also be covered by the $n_V^{\alpha,\beta}$ term.

The nature of the network used, parameters chosen and distribution of gene ranks in the network all affect which of the terms in the expression will dominate in a given scenario.

Several innovations were implemented to make the new version of the algorithm considerably faster. Firstly, instead of testing each (α, β) pair separately (identifying the regions in the original network and assessing the largest region against n_{perm} permuted

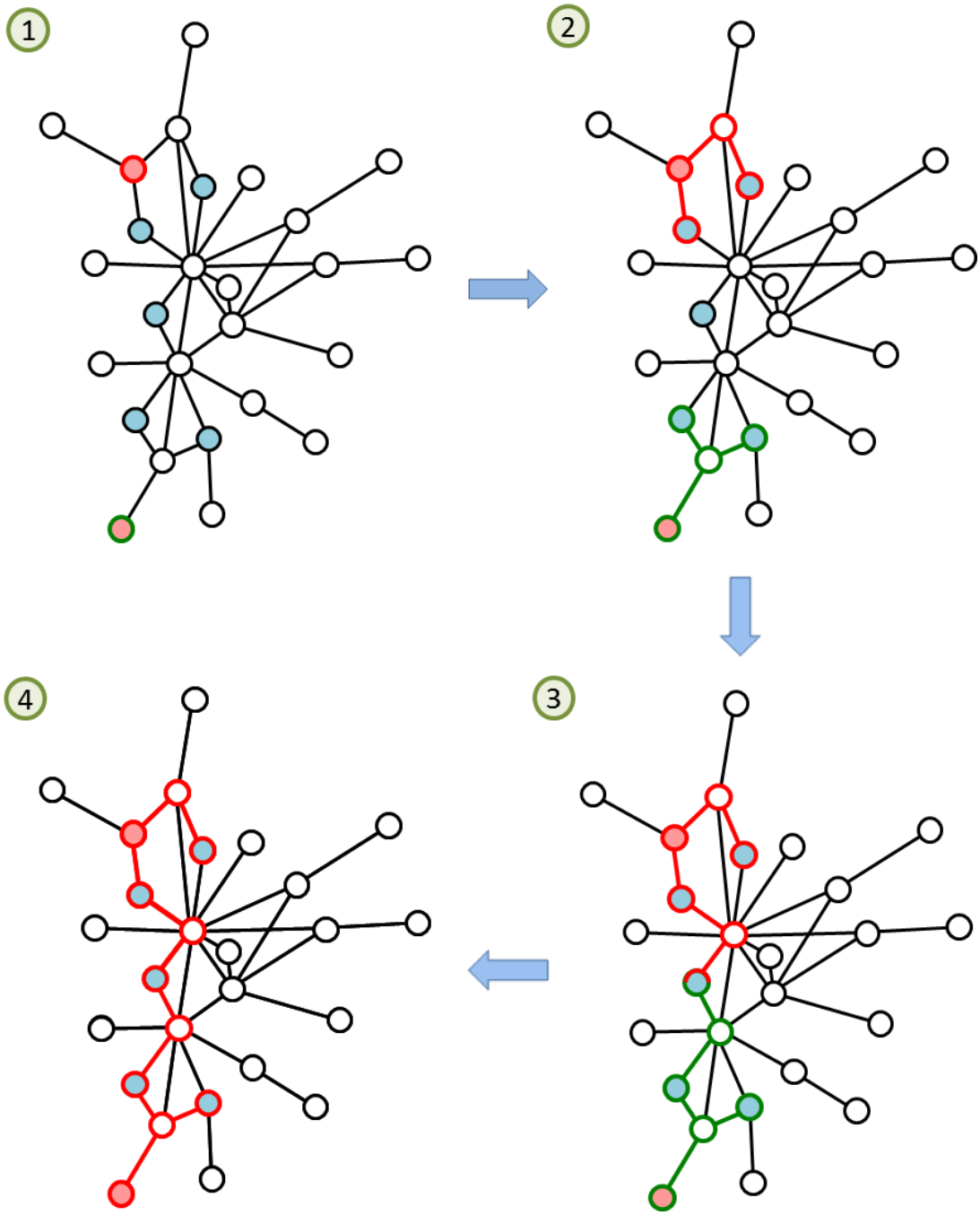


Figure 6.1 – Region Growing Analysis
Nodes in pink are seed nodes because they correspond to genes with a rank below the threshold α. In step 1, separate regions are initiated at the two seed nodes (indicated by red and green borders respectively). In steps 2 and 3 the regions are expanded to include neighbouring member nodes (with rank $<\beta$); in the case illustrated here jumps over at most one non-member node are permitted. After each iteration, regions which collide (e.g. step 3) are fused to form a single region (step 4). The process terminates at step 4 because no further member nodes can be incorporated. Figure adapted from fig. 5.1 in (Lehne 2011).

networks), all (α, β) pairs are tested together in the same set of permuted networks. This has two advantages: we need only generate n_{perm} permuted networks compared with $n_{perm}n_{\alpha\beta}$ in the original algorithm, and comparison between empirical p-values generated for different (α, β) pairs is more robust when the same permuted networks underlie the calculations.

Secondly, a nested approach is used to identify regions at different (α, β) pairs. If $\alpha \leq \beta_1 < \beta_2$ then all regions found at (α, β_1) will lie inside regions at (β_2, β_2). Therefore when performing RGA for (α, β_1) it is more efficient to search only within the regions identified for (β_2, β_2) than to search within the whole network. Likewise for $\alpha_1 < \alpha_2 \leq \beta$ all regions found at (α_1, β) lie inside regions at (α_2, β). We therefore work through the full set of (α, β) pairs ordered firstly by descending β and then by descending α.

Thirdly, a pre-processing step is performed at each (α, β) level. This removes from the network all edges that join two nodes of rank $>\beta$ (since at most one such node can be jumped in order to grow a region), and subsequently all nodes that have rank >β and degree ≤1. This can be done in $O(n_V + n_E)$ operations when pre-processing the whole network and $O\left(n_V^{\alpha,\beta} + n_E^{\alpha,\beta}\right)$ operations when pre-processing within the regions found for with thresholds α and β (because of the nested approach to testing (α, β) combinations).

Finally, for the region identification itself the improved version of RGA identifies regions sequentially using breadth-first search (Sedgewick 2003), which after pre-processing will have a running time of $O\left(n_V^{\alpha,\beta} + n_E^{\alpha,\beta}\right)$, instead of iterating over region expansion and merging steps to grow all regions simultaneously (as in the original implementation). This eliminates the need for multiple iterations.

The time-complexity of the revised algorithm is therefore approximately:

$$O\left(n_{perm}\left(n_V \log n_V + n_E + \sum_{\alpha,\beta}\left(n_V^{\alpha,\beta} + n_E^{\alpha,\beta}\right)\right)\right).$$

Since the time-complexity (∗) of the original implementation can be rewritten as:

$$O\left(n_{perm}\left(n_{\alpha\beta}(n_V \log n_V + n_E) + \sum_{\alpha,\beta} n_{iter}\left(n_V^{\alpha,\beta} + n_E^{\alpha,\beta} + m_{\alpha,\beta}\right)\right)\right),$$

where $m_{\alpha,\beta}$ is a term representing the cost of exploring the genes that neighbour the regions found using thresholds α and β, we can see that the revised version should be considerably faster due to the $n_{\alpha\beta}$-fold drop in the number of permuted networks generated, the

elimination of the n_{iter} factor, and the fact that most region-neighbouring genes are removed by pre-processing so do not need to be searched.

This performance improvement was tested by repeating an analysis from (Lehne 2011), in which regions were identified using each α in {50, 100, 150, 200, 300, 500, 1,000} (with $\beta = \alpha$) for a ranked gene list derived from the first large-scale type 1 diabetes GWAS (Wellcome Trust Case Control Consortium 2007), and compared against 10,000 network permutations. In the HuPPI2 network (described in chapter 2) this took 147 s (2.4 mins) on a standard PC using the new implementation, compared to 51,978 s (14.4 hours) using the original version with the default of 500 iterations (times are average across three executions). In theory the maximum number of iterations that could be needed is equal to the diameter of the network (the biggest minimum distance amongst all pairs of same-component genes). HuPPI2 has diameter 23, and when 23 iterations are used the original implementation takes 8,546 s (2.4 hours). The revised algorithm therefore represents a 58.1-fold improvement on this performance. In the HuPPI2_d25 network, the new implementation takes 124 s (2.1 mins) compared to 7,728 s (2.1 hours) using the original version (with 23 iterations because HuPPI2_d25 also has diameter 23). This represents a 62.4-fold improvement. As expected, the regions found by the original and new implementations match, for both networks.

The improvement in efficiency of the tool is highly advantageous because it allows analyses to cover a much wider or more granular range of α and β thresholds or to run RGA using a range of networks. This greatly improves RGA's utility as an exploratory tool.

7 Analysis of Adams-Oliver Syndrome Exome Sequence Data using Network Methods

7.1 Introduction

7.1.1 Background

Adams-Oliver syndrome (AOS; OMIM #100300) is a developmental disorder with an incidence of approximately 1 in 225,000 (Stittrich et al. 2014). It is characterised by the congenital absence of skin (aplasia cutis congenita; ACC), typically affecting the scalp, and serious limb reduction malformations (terminal transverse limb defects; TTLD) (Adams and Oliver 1945) (see Figure 7.1). However, AOS displays substantial phenotypic variability: the severity of ACC and TTLD can vary between cases to the extent that one or other may be entirely absent (in such cases a family history of the disorder indicates a positive diagnosis); additional phenotypic effects include congenital heart defects in approximately 20% of patients and a rare vascular abnormality, denoted cutis marmorata telangiectasia congenita, affecting a similar proportion of cases (Snape et al. 2009).

AOS is presumed to be a monogenic disorder. It usually occurs sporadically but can also segregate in families, where there has been evidence for both autosomal dominant (AD) and autosomal recessive (AR) modes of inheritance (Snape et al. 2009). The molecular basis of AOS is not fully characterised. Several genes have been identified as causal for AOS; however, the identification of further pathogenic variants is challenging due to both genetic heterogeneity and incomplete penetrance of disease alleles in AD kindreds (Küster et al. 1988).

Southgate *et al.* showed that two heterozygous truncating mutations (one nonsense SNV and one frameshift deletion) in the Rho GTPase-activating protein 31 gene, *ARHGAP31*, cause AOS in unrelated patients (Southgate et al. 2011). GTPases are a class of proteins that can switch between two conformational forms (active GTP-bound and inactive GDP-bound states), acting as "molecular switches" to regulate a number of cellular processes. ARHGAP31 specifically regulates Cdc42 and Rac1, which play a role in cell proliferation and migration – key aspects of organ development. Of relevance to AOS, both proteins have been shown to function during skin morphogenesis and limb development (Southgate et al. 2011). While the wild-type ARHGAP31 inactivates Cdc42 and Rac1, mutated versions of the protein were demonstrated to lead to sustained inactivation and

Figure 7.1 – Characteristic phenotype of AOS
Characteristic phenotype of AOS showing severe ACC (left panels) and a range of TTLD defects of the hands (middle panels) and feet (right panels), including partial absence of the fingers and toes and short distal phalanxes of the fingers and toes. Reprinted from (Southgate et al. 2011), with permission from Elsevier.

hence a deficiency of active Cdc42 and Rac1, indicating a gain-of-function disease mechanism for AOS (Southgate et al. 2011).

Consistent with these results, homozygous mutations in *DOCK6* have been shown to cause an AR form of AOS (Shaheen et al. 2011). While ARHGAP31 is a GTPase-activating protein, which serves to inactivate Cdc42 and Rac1 by stimulating their intrinsic GTPase activity, DOCK6 is a guanine nucleotide exchange factor which acts in the opposite direction, promoting the transition of Cdc42 and Rac1 from inactive to active states. As might be expected, loss-of-function *DOCK6* mutations, which lead to reduced levels of active GTPase, result in a similar phenotypic effect to gain-of-function *ARHGAP31* mutations (Shaheen et al. 2011).

However, in 2012 Hassed *et al.* demonstrated that an AD form of AOS could also be caused by mutations in the *RBPJ* gene, which is not known to be related to the Rho signalling family (Hassed et al. 2012). RBPJ is a transcription factor that plays a key role in the notch signalling pathway, an intercellular signalling system that is involved in a wide range of processes which include cell proliferation and differentiation during embryonic development. The heterozygous missense mutations that were shown to cause AOS impair the ability of RBPJ to bind to DNA, likely resulting in dysregulation of Notch target gene expression (Hassed et al. 2012). Notch signalling is further implicated in the aetiology of AOS by the recent discovery of causal variants in *NOTCH1* (Stittrich et al. 2014).

Finally, homozygous mutations in *EOGT* have been shown to cause an AR form of AOS (Shaheen et al. 2013; Cohen et al. 2014). The EOGT enzyme functions as a post-translational modifier of several extracellular proteins. The mechanism by which *EOGT* mutations cause AOS is not clear, but the authors suggest they could have an effect on cell-cell or cell-matrix interactions, offering a putative link to *ARHGAP31* or *DOCK6* forms of the disorder. On the other hand, EOGT has been shown to glycosylate NOTCH1 in mammals, although it is yet to be established whether a functional effect on notch signalling is seen in human cells (Shaheen et al. 2013).

The recent progress in elucidating the molecular basis of AOS has been due in large part to the availability of NGS techniques. Despite this, many cases are not explained by mutations in *ARHGAP31*, *DOCK6*, *RBPJ*, *NOTCH1* or *EOGT* (it is thought that fewer than half of AOS cases are caused by variants in one of these genes (Stittrich et al. 2014)). AOS displays a high level of genetic heterogeneity, and while there are close functional relationships between some of the implicated genes it does not appear to be a disorder of a single known functional pathway. This makes AOS an ideal candidate for network-based exome sequence analysis.

7.1.2 Analysis Strategy

Whole-exome sequencing had previously been performed for 20 individuals with AOS as part of the King's College London (KCL) rare disease programme. Of these, one resulted in the identification of *ARHGAP31* as a causal gene for AOS (Southgate et al. 2011) and two more (independently of the first published report) in the identification of *NOTCH1* (Southgate *et al*., unpublished results). Three further exomes harbour variants which are presumed to explain the occurrence of the disease for these individuals: two exomes carry compound heterozygous mutations in *DOCK6* and one a heterozygous missense mutation in *RBPJ*. However, progress using a simple intersection filtering approach has been limited and for the remaining 14 exomes the disease-causing variants have not been identified up to this point. These exomes will subsequently be referred to as the *unsolved cases*.

Since locus heterogeneity is an established feature of AOS, a network-based approach to identifying further causal genes in this cohort is justified. The aim of such methods, to find connected sets of genes that represent functional pathways underlying AOS (thereby helping to illuminate the disease mechanism), is reasonable: among the known AOS genes, *ARHGAP31* and *DOCK6* are functionally related, as are *RBPJ* and *NOTCH1*. To date, however, no network-based analyses of AOS have been reported.

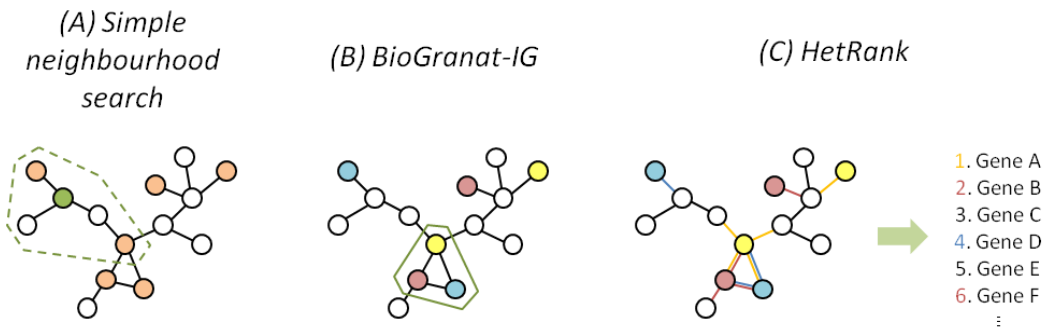


Figure 7.2 – Network-based methods used to analyse AOS exome data
(A) Simple neighbourhood search: for this candidate-gene approach the neighbourhoods of known AOS genes (green nodes) are examined for overrepresentation of post-filtering variants in the sample exomes (orange nodes). (B) BioGranat-IG: red, blue and yellow nodes signify post-filtering variants in different sample exomes. The tool seeks a small connected set of genes in which all or most sample exomes are mutated. (C) HetRank: red, blue and yellow nodes as before. Genes are ranked according to the likelihood that they cause AOS based on characteristics of the variants they contain and variants in neighbouring genes in complementary sample exomes.

Three independent network methods were applied to the exome data to propose new AOS genes. Firstly a candidate-gene approach was taken, with the network neighbourhoods of known AOS genes being examined for overrepresentation of post-filtering (that is, rare and non-synonymous) sequence variants. The other approaches were hypothesis-free (they did not rely on the known AOS genes). The BioGranat-IG tool developed in chapter 4 was used to search for subnetworks in which all or most exomes carry a post-filtering variant. Finally, the HetRank tool developed in chapter 5 was used to generate a list of genes prioritised with respect to evidence for disease causality. The different approaches are illustrated in Figure 7.2.

Several interaction networks were used to perform the analyses, including three protein interaction networks (PINs), a co-expression network and an integrated network of several interaction types. Use of different networks allows different types of functional pathways to be examined as potentially underlying AOS. Additionally the BioGranat-IG and HetRank analyses were performed separately for the full set of sample exomes and for the subset of unsolved cases. In the former case this allows us to observe whether the methods prioritise the known AOS genes, while in the latter case any confounding information from the non-causal variants in the solved cases is removed. (For BioGranat-IG an implicit candidate-gene approach could also be explored by using the full set of exomes but filtering out all but the causal variants in the solved cases.)

KGGSeq-prioritisation and Region Growing Analysis (RGA), described fully in the previous chapter, were used for the interpretation of BioGranat-IG and HetRank results, respectively.

7.2 Methods

7.2.1 Exome Data

Whole exome sequencing was performed on 20 individuals with AOS and 336 unaffected controls according to the procedure described in chapter 2, section 2.3.

Table 7.1 lists the properties of the 20 sequenced AOS exomes. All affected individuals have a primary diagnosis of AOS but there is some variation in clinical phenotypes.

At 20× coverage, exome capture ranged from 71.0% to 92.6%; while 80-90% is considered to be a normal range of coverage, 90% or higher is desirable to maximise the power to accurately detect sequence variants (Mertes et al. 2011). For six of the affected individuals causal mutations were previously identified. For one of these (exome S0334) one of the two variants in *DOCK6* that make up the causal compound heterozygous mutation fell outside the captured region and had been subsequently identified by Sanger sequencing. Therefore the exome data used for this analysis does not include this variant.

By considering the family histories of the affected individuals, 15 of the 20 were assumed to have an AD form of AOS, three an AR form, one either AD or AR, and one either AD or X-linked.

Two of the affected individuals (S0039 and S0301) were related. Since AOS is an inherited disorder it is expected that the same mutation causes AOS in both family members. Therefore for all analyses these two exomes (having 18,915 and 23,964 called variants respectively) were replaced by a single file comprising only the 10,831 called variants shared by both individuals, effectively reducing the dataset to 19 AOS exomes (of which 13 are unsolved cases). All other affected individuals were unrelated.

The 336 unaffected controls (a subset of the exomes covered by the in-house exome database which will be referred to subsequently as *non-AOS control exomes*) represent unrelated individuals of European ancestry sequenced as part of the KCL rare disease programme. As such they may include causal variants for a range of diseases other than AOS. Since these diseases are clinically diverse the data are considered suitable for use as controls for both variant filtering (see section 7.2.2) and for use by the HetRank tool (section 7.2.6).

7.2.2 Variant Filtering

To perform the simple neighbourhood search and BioGranat-IG analyses, variants in the exomes of affected individuals were filtered to exclude from consideration those less

Table 7.1 – Properties of AOS exomes
Blue shading indicates that causal variant has already been identified. Yellow shading denotes exomes S0039 and S0301 derived from related AOS patients. CHD = congenital heart defect. * = one of the two variants in *DOCK6* that make up the causal compound heterozygous mutation for exome S0334 was not captured by whole exome sequencing.

Exome ID	Mode of inheritance	Diagnosis	Exome coverage (to 20×)	Causal mutation(s) identified	Notes
S0038	AD	AOS	71.4%	*ARHGAP31*	-
S0039	AD	AOS	71.2%	-	Related to S0301
S0040	AD	AOS	71.0%	-	-
S0069	AD	AOS	80.9%	-	-
S0301	AD	AOS	84.5%	-	Related to S0039
S0302	AD	AOS	76.6%	-	-
S0304	AD	AOS	83.2%	-	-
S0305	AR	AOS	81.7%	-	Parents 1st cousins
S0306	AD	AOS	86.1%	-	-
S0307	AD	AOS	84.8%	-	-
S0308	AD or AR	AOS	92.6%	-	-
S0311	AD	AOS	82.5%	*NOTCH1*	-
S0332	AR	AOS / Orstavik syndrome	85.1%	*DOCK6*	-
S0333	AD	AOS	83.2%	-	-
S0334	AR	AOS / Orstavik syndrome	85.9%	*DOCK6**	-
S0335	AD	AOS	79.2%	*NOTCH1*	-
S0336	AD	Probable AOS	80.0%	-	-
S0337	AD	AOS	77.6%	*RBPJ*	-
S0338	AD or X-linked	AOS / Dandy-Walker syndrome / CHD	86.3%	-	-
S0339	AD	AOS / Sorsby syndrome	90.2%	-	-

likely to cause a rare monogenic disease. Five levels of filtering were applied, as illustrated in Figure 7.3. For notational convenience these are termed filtering levels 1-5.

At level 1, all 19 exomes underwent the same filtering steps (no special treatment for the six solved cases with known causal variants). Synonymous variants were excluded. No assumptions were made about the mode of inheritance of AOS for each exome, so that three mutation types were included: heterozygous (AD inheritance), homozygous (AR) and compound heterozygous (AR). Since AOS is a very rare disorder, only novel and very rare variants were retained. Heterozygous variants were required to have alternative allele frequencies of <0.1% according to both 1000 Genomes Project and Exome Variant Server

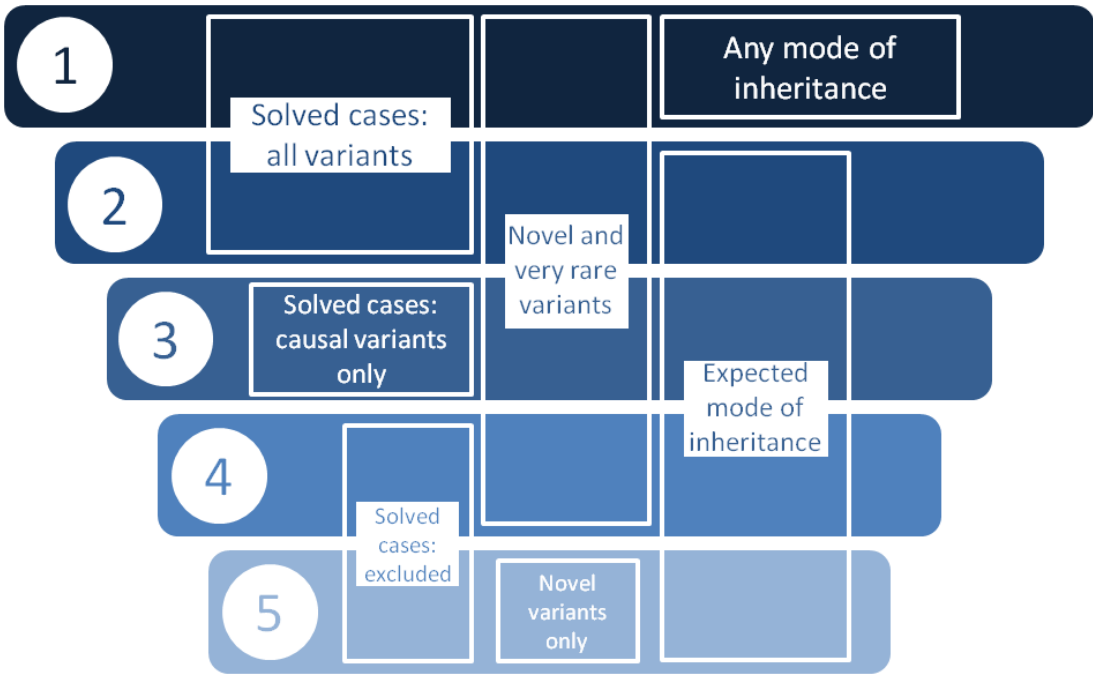


Figure 7.3 – Variant filtering levels for AOS exomes
Filtering levels differ according to: treatment of solved cases (treat same as unsolved cases, include only the true causal variant(s) or remove from analysis entirely); frequency threshold (allow novel variants only, or also include known variants with sufficiently low minor allele frequency as described in main text); and mode of inheritance (make no assumption, or assume variant will be consistent with expected mode of inheritance).

(EVS) annotation, and to have been observed once or fewer in heterozygous form (and never in homozygous form) in the in-house exome database (approximately 850 exomes in total). Homozygous variants were required to have alternative allele frequencies of $<\sqrt{0.1\%}$ and one or fewer previous observations in homozygous form in the in-house exome database. Compound heterozygous variants were included if two or more heterozygous variants in the same gene had alternative allele frequencies of $<\sqrt{0.1\%}$ and one or fewer previous observations in homozygous form in the in-house exome database. Finally, variants were filtered at the gene level against the 336 non-AOS control exomes in order to address the problem that large and highly-polymorphic genes are more likely to be falsely implicated by variant-filtering approaches (as discussed in section 1.3.5). Variants were excluded in any gene in which 20 or more non-AOS control exomes, after being subject to the same filtering steps, carried a post-filtering variant.

At level 2, variants were additionally filtered to include only those matching the expected mode of inheritance for each AOS exome. The individuals S0332 and S0334 were assumed to have an AR form of AOS; only homozygous and compound heterozygous variants were included for these exomes. Exome S0305 was obtained from an individual whose parents are first cousins; for this exome only homozygous variants were included.

Individual S0338 (male) was assumed to have either an AD or X-linked form of AOS; a conservative approach was taken by including heterozygous variants from all chromosomes as well as homozygous and compound heterozygous X chromosome variants. The mode of inheritance for exome S0308 had not been determined and therefore no additional filtering on variant zygosity was performed. All other cases were assumed to have an AD inheritance pattern hence only heterozygous variants were included. Filtering thresholds were otherwise identical to level 1. Gene-wise filtering against non-AOS control exomes was performed separately for each mode of inheritance using the appropriate filtering for controls.

Level 3 provided an implicit candidate-gene approach. At level 3, filtering criteria were identical to level 2 for the 13 exomes representing unsolved cases but for the six solved cases only the true causal variants were included. In the case of exome S0334, for which AOS is caused by a compound heterozygous mutation in *DOCK6*, one of the true causal variants was not captured by the whole exome sequencing process. This meant *DOCK6* variants for this exome had been excluded at filtering level 2; the missing variant was artificially added at level 3 to ensure that each solved case was represented by variant(s) in the appropriate gene. Likewise, for exomes S0311 and S0335, variants in *NOTCH1* had been excluded at filtering level 2 because 20 of the non-AOS control exomes carried a variant in that gene, therefore at filtering level 3 these variants were manually added back. For all other solved cases no manual re-instatement of variants was necessary.

At level 4, the exomes from solved cases were removed from the analysis entirely, permitting an analysis of the unsolved cases that was unbiased by the previously-solved cases. All filtering criteria for the 13 unsolved cases were as at level 3.

Finally, filtering at level 5 was identical to level 4 except that previously-observed variants were excluded to leave only novel variants. In practice, any variant annotated with an alternative allele frequency (of any value) from EVS or 1000 Genomes Project data, or contained in the in-house exome database in homozygous or heterozygous form, was excluded. As with the other filtering levels, gene-wise filtering against controls used matching criteria: variants in any gene in which 20 or more of the 336 non-AOS control exomes carried a novel (non-synonymous) variant were excluded.

Table 7.2 lists the number of called variants for each exome at each level of filtering.

7.2.3 Interaction Networks

Analyses were performed using several of the networks described in detail in chapter 2. These comprised three PINs (PINA and the two smaller but higher-confidence

Table 7.2 – Called variants in AOS exomes at each level of filtering
* = for exome S0334 one of the true causal variants in gene *DOCK6* was not captured by the whole exome sequencing process; † = for exomes S0311 and S0335 variants in *NOTCH1* would ordinarily be filtered out because more than 20 non-AOS controls carry a post-filtering variant in this gene. At filtering level 3 appropriate variants were artificially added for these exomes to ensure that each solved case was represented by a variant in its causal gene.

Exome ID	Number of called variants					
	Unfiltered	Filtering level 1	Filtering level 2	Filtering level 3	Filtering level 4	Filtering level 5
S0038	19,277	233	180	1	0	0
S0039_S0301	10,831	85	65	65	65	31
S0040	18,721	243	195	195	195	119
S0069	21,363	223	165	165	165	94
S0302	22,877	256	195	195	195	113
S0304	23,575	258	220	220	220	146
S0305	24,606	475	69	69	69	4
S0306	23,762	201	162	162	162	84
S0307	23,797	200	164	164	164	86
S0308	24,839	236	236	236	236	115
S0311	23,326	224	171	1†	0	0
S0332	23,911	250	112	2	0	0
S0333	23,613	238	185	185	185	107
S0334	23,685	221	100	2*	0	0
S0335	23,772	286	243	1†	0	0
S0336	22,965	196	159	159	159	91
S0337	22,930	212	166	1	0	0
S0338	25,638	451	281	281	281	168
S0339	24,459	225	178	178	178	96

networks PINAmin2 and CPDBconf95), a co-expression network (COXPRES30) and a network integrating several types of interaction data (Multinet).

Since hub genes can be overrepresented in BioGranat-IG results, the hub-free networks PINA_d50, PINAmin2_d50, CPDBconf95_d50, COEXPRES30_d50 and Multinet_d50 (described in section 2.1.2) were used for BioGranat-IG analyses. All other analyses use the full interaction networks.

To ensure consistency between network node labels and exome gene symbols, a list of gene symbol synonyms was obtained from the HUGO Gene Nomenclature Committee (Gray et al. 2013). An R programme was implemented to replace the node label with a synonymous symbol for any network node whose label did not map to any gene in the exome data, but for which an unambiguous mapping existed for one of the synonyms. However, in practice no changes were necessary for any of the networks.

7.2.4 Simple Neighbourhood Search

Since several AOS genes are known, a logical first step in seeking causal variants for the 13 unsolved cases was to examine variants in genes which directly or indirectly interact with known AOS genes. The following procedure was employed for this purpose.

Given a set of genes X, two test statistics can be defined to assess the likelihood of genes in X being involved in AOS:

- $S_1(X)$, being the number of the 13 unsolved cases in which a post-filtering sequence variant is observed in any gene in X, and
- $S_2(X)$, being the number of genes in X which contain a post-filtering sequence variant for two or more of the 13 unsolved cases.

For an undirected network G and seed gene g in G, let $N_d(g)$ denote the neighbourhood comprising the set of all genes in G reachable from g via d interactions or fewer. Evidence that variants in genes in $N_d(g)$ may cause AOS in the unsolved cases can be quantified by comparing $S_1(N_d(g))$ and $S_2(N_d(g))$ against $S_1(X_k)$ and $S_2(X_k)$, k = 1, ..., 10,000, where each X_k is a set of genes of size $|N_d(g)|$ selected uniformly randomly from all genes in G.

This process was carried out for all five networks. For each network the process was repeated using each known AOS gene as the seed gene (provided it was present in the network – see results section 7.3.1 below), and additionally using all known AOS genes as seed genes simultaneously. Neighbourhood parameters d = 1 and 2 were both tested. Since the gene *UBC* (one of the genes which produces ubiquitin) is exceptionally highly connected in several of the networks, and is the node of highest degree in PINA, PINAmin2, CPDBconf95 and Multinet, additional tests were performed in these networks using the modified neighbourhoods $N_2^*(g)$, comprising all genes in G reachable from seed gene g via two interactions or fewer but excluding paths via the node representing *UBC* (and *UBC* itself). The process was carried out separately for variants subject to level 4 filtering and to level 5 filtering.

Neighbourhood identification was performed using UNIX commands and permutation testing was performed in R.

7.2.5 BioGranat-IG Analysis

BioGranat-IG analysis, as detailed in chapter 4, was performed for all five (hub-free) networks with variants filtered at each of levels 1-4.

For each network and filtering level, several BioGranat-IG analyses were performed to allow a thorough interrogation of the exome data. Firstly, (exact) triplet and quadruplet searches were performed to examine how many AOS cases could potentially be explained

by three or four interacting genes. These were repeated for "optimal" (size flexibility = 0; number flexibility = 0) and "near-optimal" (size flexibility = 1; number flexibility = 1) parameter sets. Secondly combined (heuristic) minimum distance and multi-minimum distance searches were performed to examine how few genes could potentially be sufficient to explain all (or most) AOS cases. These were repeated with the searches limited to subnetworks of size ten or smaller, and with the searches not limited by subnetwork size. Only optimal results were considered for these runs.

To facilitate interpretation of subnetworks found by BioGranat-IG, KGGSeq-prioritisation (as described in chapter 6) was used to prioritise results.

7.2.6 HetRank Analysis

HetRank analysis, as detailed in chapter 5, was performed for all five networks. For each network, separate analyses were performed on the set of all 19 exomes and the set comprising only the 13 unsolved cases.

Ranking factors consisted of zygosity, variant consequence (e.g. "nonsynonymous SNV", "stopgain SNV", "frameshift deletion"), EVS and 1000 Genomes Project alternative allele frequencies (two factors) and number of observations of the variant in the in-house exome database in homozygous and heterozygous form (two factors). Since the majority of AOS cases in this study are assumed to have AD inheritance, the zygosity ranking places heterozygous variants ahead of homozygous variants. Since AOS represents a severe phenotypic consequence for a single mutation, the variant consequence ranking places nonsense ("stopgain") SNVs and frameshift indels/substitutions ahead of all other non-synonymous variants, which in turn rank ahead of synonymous SNVs. The other factors are ranked numerically in increasing order.

In all analyses the full set of 336 non-AOS control exomes were used to generate a final ranking factor to down-weight genes which rank highly in unaffected controls.

Unless otherwise stated, the weight parameters used were those determined in chapter 5, section 5.2.4, to give an optimal prioritisation based on test data. Thus weights of 1 were used for all ranking factors except for the number of observations of the variant in the in-house exome database in homozygous form, and the factor derived from non-AOS control exomes, both of which were given a weight of 8.

Finally, RGA (described fully in chapter 6, section 6.2) is used to analyse the prioritised gene lists returned by HetRank, in the same network that HetRank has used for ranking. In each case all thresholds in the range $1 \leq \alpha \leq 250$ are tested, with $\beta = \alpha$. "Jumps" are not permitted (because HetRank itself can incorporate information from indirectly-connected genes to produce its rankings). 10,000 degree-constrained permuted networks are

used to estimate significance, with RGA's default standard deviation of 5.0 nodes used for label-shuffling.

7.2.7 Tools Used for Analysis of Results

Network diagrams are generated using Cytoscape (Smoot et al. 2011).

Existing functional annotation for subnetworks or gene lists of interest is explored using Gene Ontology (GO) term enrichment analysis. GO terms provide a controlled vocabulary describing biological function (Ashburner et al. 2000). GO terms form a directed acyclic graph, which means that a term *t* (e.g. "endocardium development") can have parent terms (describing more general functions of which *t* is a type, e.g. "heart development") and child terms (more specific functions that are examples of *t*, e.g. "endocardium morphogenesis"). All genes annotated with function *t* are necessarily annotated with all parent terms of *t*. Enrichment testing examines whether any terms are overrepresented among a specific set of genes, relative to a background gene-set from which these genes were selected.

Enrichment testing is performed with the Web-based Gene Set Analysis Toolkit (WebGestalt) (Wang et al. 2013a), using default parameters. For subnetworks of interest the background gene set comprises all genes in the network from which the subnetwork was extracted; for gene lists of interest generated by HetRank the background gene set comprises all genes assigned a final rank (that is, all network genes plus any other gene containing a variant in any case exome). Unless otherwise stated, only GO biological process terms are tested. Enrichment p-values after adjustment for multiple testing (using Benjamini and Hochberg's method which controls the false discovery rate (Benjamini and Hochberg 1995)) are presented using the notation *adjP*.

Unless otherwise referenced, summaries of individual gene function are obtained from GeneCards (www.genecards.org, Stelzer et al. 2011).

7.3 Results and Discussion

7.3.1 Simple Neighbourhood Search

For each network Table 7.3 lists the number of direct (exactly one interaction distant) and indirect (exactly two interactions distant) neighbours for each known AOS gene. Gene *EOGT* is not present in any of the networks and therefore was not analysed. *DOCK6* is not present in the network CPDBconf95 and is only connected to the extreme hub gene *UBC* in PINAmin2. Otherwise all AOS genes were analysed in all networks.

Table 7.3 – Properties of known AOS genes in interaction networks
"Direct neighbours" are distance 1 from known AOS gene; "Indirect neighbours" are distance 2; "Indirect neighbours (excluding *UBC*)" excludes *UBC* and genes reached exclusively via *UBC*. In the COXPRES30 network *UBC* is not an extreme hub gene so no entries are given in column "Indirect neighbours (excluding *UBC*)". Table continues onto next page.

Network	Known AOS gene	In network?	Direct neighbours	Indirect neighbours	Indirect neighbours (excluding *UBC*)	Notes
PINA	*ARHGAP31*	✓	5	560	559	
	DOCK6	✓	2	7,819	102	*NOTCH1* is an indirect neighbour via *UBC*
	EOGT	✗	-	-	-	
	NOTCH1	✓	65	8,613	2,596	*RBPJ* is a direct neighbour *DOCK6* is an indirect neighbour via *UBC*
	RBPJ	✓	17	827	826	*NOTCH1* is a direct neighbour
PINAmin2	*ARHGAP31*	✓	2	42	41	
	DOCK6	✓	1	4,777	0	*NOTCH1* is an indirect neighbour via *UBC*
	EOGT	✗	-	-	-	
	NOTCH1	✓	13	4,829	204	*RBPJ* is a direct neighbour *DOCK6* is an indirect neighbour via *UBC*
	RBPJ	✓	2	51	50	*NOTCH1* is a direct neighbour
CPDBconf95	*ARHGAP31*	✓	2	43	43	
	DOCK6	✗	-	-	-	
	EOGT	✗	-	-	-	
	NOTCH1	✓	14	951	196	*RBPJ* is a direct neighbour
	RBPJ	✓	16	586	585	*NOTCH1* is a direct neighbour

Table 7.3 – Properties of known AOS genes in interaction networks (continued)

Network	Known AOS gene	In network?	Direct neighbours	Indirect neighbours	Indirect neighbours (excluding *UBC*)	Notes
COXPRES30	*ARHGAP31*	✓	7	72	-	
	DOCK6	✓	4	54	-	
	EOGT	✗	-	-	-	
	NOTCH1	✓	3	54	-	
	RBPJ	✓	5	190	-	
Multinet	*ARHGAP31*	✓	1	91	90	
	DOCK6	✓	1	104	104	
	EOGT	✗	-	-	-	
	NOTCH1	✓	36	3,305	2,320	*RBPJ* is a direct neighbour
	RBPJ	✓	22	2,482	2,481	*NOTCH1* is a direct neighbour

Table 7.4 – AOS Simple neighbourhood search: permutation tests with significant results
Neighbourhood: plus2 = genes within distance 2 of starting gene (d = 2); plus2_noUBC = genes within distance 2 of starting gene excluding *UBC* and genes reached via *UBC*. Variant type: novel_rare = subject to level 4 variant filtering; novel = subject to level 5 variant filtering. Empirical p-values: proportion of 10,000 random tests in which a greater of equal test statistic was observed; nominally significant (<0.05) p-values are highlighted in green.

Starting gene	Network	Neighbourhood	Variant type	Genes in neighbour-hood	Observed AOS cases covered (S_1)	S_1 empirical p-value	Observed multiply-mutated neighbours (S_2)	S_2 empirical p-value
ARHGAP31	PINA	plus2	novel	566	13	0.0310	2	0.5665
ARHGAP31	PINA	plus2_noUBC	novel	565	13	0.0278	2	0.5731
ARHGAP31	PINAmin2	plus2	novel_rare	45	9	0.0073	1	0.3335
ARHGAP31	PINAmin2	plus2_noUBC	novel_rare	44	9	0.0057	1	0.3330
ARHGAP31	Multinet	plus2	novel	93	8	0.0328	1	0.2158
ARHGAP31	Multinet	plus2	novel_rare	93	12	0.0019	3	0.0446
ARHGAP31	Multinet	plus2_noUBC	novel	92	8	0.0333	1	0.2086
ARHGAP31	Multinet	plus2_noUBC	novel_rare	92	12	0.0018	3	0.0443
RBPJ	COXPRES30	plus2	novel	196	11	0.0188	2	0.1204
All AOS genes	COXPRES30	plus2	novel_rare	383	12	0.6978	7	0.0483

104 different combinations of interaction network, AOS gene (where present in a network), neighbourhood type ($d = 1$, $d = 2$ or $d = 2$ excluding *UBC*) and variant filtering level (4 = very rare and novel variants; 5 = novel variants only) were tested using both test statistics, S_1 (number of the 13 unsolved AOS cases with a variant in the neighbourhood) and S_2 (number of genes in the neighbourhood with variants in two or more unsolved AOS cases).

In addition 28 different combinations of interaction network, neighbourhood type and variant filtering level were tested using all known AOS genes simultaneously as seed genes.

In the 104 tests using individual AOS genes, the value of test statistic S_1 ranged from 0 to all 13 unsolved cases having a variant in the neighbourhood and the value of test statistic S_2 ranged from 0 to 90 genes in the neighbourhood containing variants for multiple AOS cases. However, since the neighbourhoods tested vary widely in size from 1 to 8,613 genes the test statistic values should be interpreted using the permutation test, which considers how the gene neighbourhoods compare to 10,000 randomly selected gene sets of the same size.

Table 7.4 presents the nine tests in which at least one of the test statistics S_1 and S_2 had a nominally significant empirical p-value at the 5% level. None of the tests considering direct neighbourhoods ($d = 1$) were significant.

Eight tests considering the AOS gene *ARHGAP31* were significant. These reduce to four scenarios since *ARHGAP31* is not directly connected to *UBC* in any of the networks and hence the indirect ($d = 2$) and "indirect excluding *UBC*" neighbourhoods provide equivalent results for the purposes of this analysis (and only the indirect neighbourhoods will be examined here). Empirical distributions based on the 10,000 random tests are shown in Figure 7.4 for these four scenarios.

The smallest p-value was observed for the S_1 statistic based on novel and rare variants in the Multinet network ($p = 0.0019$), and this was the only scenario which also had a significant p-value for the S_2 test ($p = 0.0446$). In a neighbourhood of 93 genes, 12 of the 13 unsolved AOS cases carried a very rare or novel variant in at least one of these genes, and three of these genes harboured a variant for two or more of the AOS exomes. The p-values suggest that these observations are higher than would typically be expected for a randomly-selected set of 93 genes, implying an enrichment of rare or novel variation. It is therefore reasonable to examine the variants which fall in the neighbourhood; while a certain proportion of these could be due to chance alone (as illustrated by Figure 7.4a) the significant enrichment suggests that at least some of the variation could be linked to AOS.

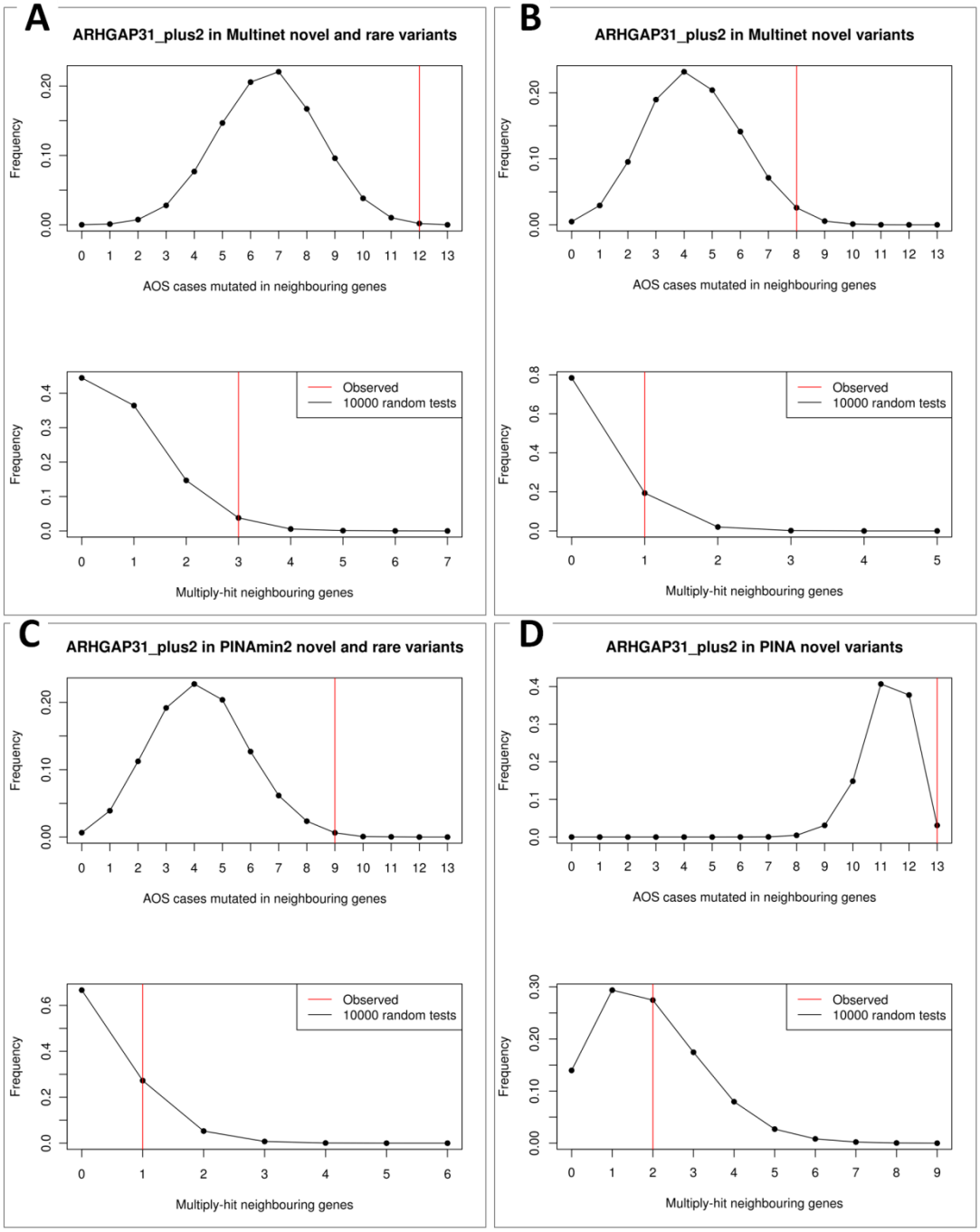


Figure 7.4 – Empirical distributions for *ARHGAP31* neighbourhoods with significant test statistics
In each panel, top graph shows empirical distribution of statistic S_1 (unsolved AOS cases carrying a variant) and bottom graph shows distribution of statistic S_2 (genes containing a variant in two or more unsolved cases) in 10,000 random tests based on neighbourhood size. (a) network = Multinet, neighbourhood distance $d = 2$, variant type = novel and rare, S_1 and S_2 are both significant; (b) network = Multinet, $d = 2$, variant type = novel only, S_1 is significant; (c) network = PINAmin2, $d = 2$, variant type = novel and rare, S_1 is significant; (d) network = PINA, $d = 2$, variant type = novel only, S_1 is significant.

The three genes in which two AOS exomes harbour a variant are *MAGI1*, *MAP3K11* and *SYNJ2*. (Note that exome S0040 carries a variant in both *MAGI1* and *SYNJ2*, which could imply that one of these is not causal for AOS since only one causal variant is expected per exome – this assumes that the variants do not act additively through a digenic disease mechanism.) In total 14 other genes contain variants. These are *CIT*, *CNTNAP1*, *DOCK1*, *DOCK2*, *IL1RAP*, *KALRN*, *LATS1*, *MAP3K4*, *MCF2L*, *MCM3AP*, *MYH9*, *PARK2*, *SH3RF1* and *TIAM1*. The logic behind the simple neighbourhood search is to start at a known AOS gene and search for directly or indirectly interacting genes that may have a disease role. Therefore further consideration of this gene list should take into account known gene function: would a variant in any of these genes impact the regulation of Rho GTPases, as AOS-causing variants in the seed gene *ARHGAP31* do? Gene function will be considered further below.

A significant p-value was also observed in the same network (Multinet) for the S_1 statistic based on novel variants only ($p = 0.0328$; see Figure 7.4b). In this case 8 of 13 AOS exomes had a novel variant in the 93 genes in the neighbourhood. Genes with novel variants are *CIT*, *CNTNAP1*, *DOCK2*, *IL1RAP*, *KALRN*, *MAP3K11* (in two exomes, although the S_2 p-value was not significant in this test), *MAP3K4*, *MCF2L*, *MYH9*, *PARK2* and *SYNJ2*. If one of the original assumptions of intersection filtering (discussed in section 1.3.4 of the thesis introduction), that a highly-penetrant disease-causing variant is unlikely to have been listed in a database of sequence variation without linking it to the disease phenotype, is used, it could be argued that this subset of genes should be prioritised for follow-up study.

In the PINAmin2 network, nine AOS cases carry novel or rare variants in a neighbourhood of 45 genes (S_1 p-value 0.0073; Figure 7.4c). Genes that contain variants are *CLTC*, *DNM1*, *DOCK1*, *DOCK2*, *ITSN1*, *MCF2L*, *SH3RF1*, *SOS1*, *SYNJ2* (variants in two exomes but S_2 p-value not significant) and *TIAM1*. In the PINA network all 13 AOS cases carry novel variants in a neighbourhood of 566 genes (S_1 p-value 0.0310; Figure 7.4d). While we see a significant empirical p-value the fact that we observe 35 variants across 33 of the 566 genes makes these results less informative than those observed in Multinet and PINAmin2 so will not be discussed further.

In total this approach has highlighted 21 genes (17 from Multinet and 10 from PINAmin2, with 6 common to both networks) having post-filtering variants in genes that interact with *ARHGAP31*, either directly or indirectly. Follow-up study would be needed to confirm that any of these variants cause AOS, which could include laboratory-based experiments such as screening for variants in a larger cohort of AOS patients or performing functional studies in cell lines or animal models to demonstrate that the AOS phenotype can feasibly result from these variants. Due to the expensive and time-consuming nature of such

experiments a more compelling argument for these variants' involvement is required. To this end we might look particularly at genes with variants in multiple AOS cases, novel variants, variants predicted to have a more severe effect on the protein product, and genes where there is some existing functional knowledge suggesting a plausible disease mechanism consistent with that of *ARHGAP31*.

Four of the genes feature (along with *ARHGAP31*) in the Reactome curated gene set "Signalling by Rho GTPases" (downloaded from the Molecular Signatures Database [MSigDB] on 30th April 2014 (Subramanian et al. 2005)). Novel missense SNVs were found in *ITSN1* (exome S0336) and *KALRN* (exome S0308), and very rare missense SNVs were found in *SOS1* (exome S0305) and *TIAM1* (exome S0308). It should be noted, however, that one of the reasons for using interaction networks for this type of analysis is to exploit experimentally-obtained evidence for functional relationships that have not yet been fully understood and added into curated pathways (Lehne and Schlitt 2012). Therefore genes outside of this curated pathway are also of interest. Notably *MAP3K11* harbours a novel nonsense SNV in exome S0069, as well as a novel missense SNV in S0304. *MAP3K11* is a kinase in the MAPK signalling pathway, known to have a number of roles including involvement "in the transcription activity of NF-kappaB mediated by Rho family GTPases and CDC42" (www.genecards.org, Stelzer et al. 2011). In addition, novel splice-site mutations were found in *CLTC*, *IL1RAP*, *MCF2L* and *PARK2* while a novel (non-frameshift) insertion was found in *MAP3K4*, and these genes may also be worth investigating further.

Besides *ARHGAP31*, *RBPJ* was the only other known AOS gene for which a significant p-value was observed. In the COXPRES30 network, 11 of the unsolved AOS cases carried a novel variant in the $d = 2$ neighbourhood of 196 genes around *RBPJ* (S_1 p-value = 0.0188, see Figure 7.5). Variants were found in the 15 genes *BMP2K*, *DCP1A*, *DENND4C*, *EP400*, *JMJD1C*, *KIAA1109*, *NKTR*, *NR1D2*, *PUM1*, *RIF1*, *ROCK1*, *SMC5*, *UBR1*, *VPS13B* and *YTHDC1*. Interpretation of this gene set is more difficult than for the results discussed in the previous paragraphs. One reason for this is that the precise nature of the interactions in the COXPRES30 network, where genes are connected if their expression profiles across a range of microarray samples are sufficiently similar, is less clear than the physical interactions represented in the PINAmin2 and Multinet networks. Since mutations in *RBPJ* are thought to cause AOS by disrupting the notch signalling process we can check this gene list for genes relevant to this process. However, none are present in the KEGG "Notch signalling pathway" obtained via MSigDB (Subramanian et al. 2005). In order to explore any shared function among the 15 genes found, a GO term enrichment analysis was performed; however, no significant enrichment for any GO biological process term was

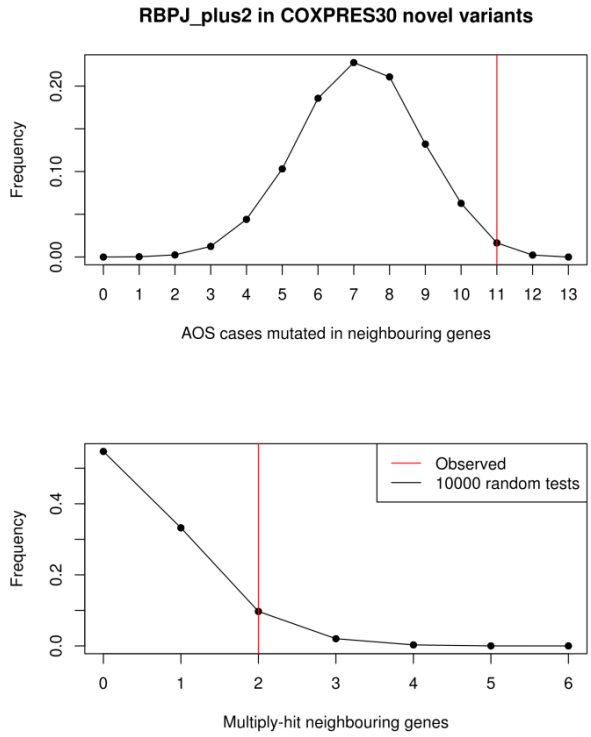


Figure 7.5 – Empirical distributions for novel variants in *d* = 2 neighbourhood of *RBPJ* in COXPRES30
Top graph shows empirical distribution of statistic S_1 (unsolved AOS cases carrying a variant) and bottom graph shows distribution of statistic S_2 (genes containing a variant in two or more unsolved cases) in 10,000 random tests based on neighbourhood size. S_1 is significant.

observed (data not shown). A handful of the genes stand out due to the type of variants present. In particular, *BMP2K* contains a novel splice-site variant for the shared S0039/S0301 exome and a novel (non-frameshift) insertion for exome S0306, while *EP400* has both a novel (non-frameshift) deletion and a novel missense SNV in exome S0307 (as well as a novel missense SNV in exome S0339). Novel splice-site variants were also observed in *KIAA1109* and *NKTR*. However, since the genes found in PINAmin2 and Multinet's *ARHGAP31* neighbourhoods have better characterised functional links, it would be imprudent to recommend follow-up analysis of any of the genes found in COXPRES30's *RBPJ* neighbourhood ahead of those.

Finally it is worth noting that of the 28 tests that used all five AOS genes simultaneously as seed genes, only one produced a significant result. In the COXPRES30 network, more genes than would be expected by chance harbour novel or very rare variants in two or more AOS exomes, and are within distance $d = 2$ of a known AOS gene ($S_2 = 7$, $p = 0.0483$; see Figure 7.6). These genes are *BMP2K*, *JMJD1C*, *KIAA1109* and *RANBP2* which interact indirectly with *RBPJ*; *COL4A2* (indirect interaction with *DOCK6*); *PCDH12* (indirect interaction with *ARHGAP31*), and *SBF1* (indirect interaction with *NOTCH1*). (Note that we previously saw novel variants in two exomes in the gene *EP400*, an indirect

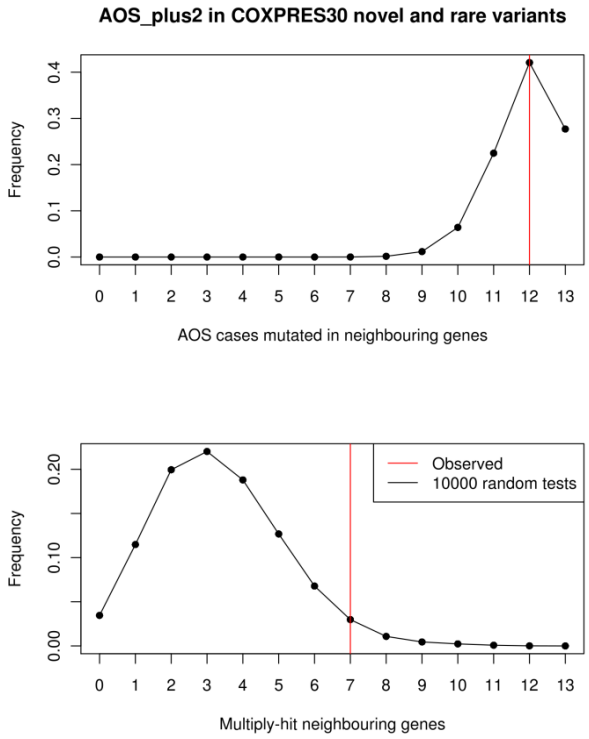


Figure 7.6 – Empirical distributions for novel and very rare variants in combined *d* = 2 neighbourhoods of all known AOS genes in COXPRES30
Top graph shows empirical distribution of statistic S_1 (unsolved AOS cases carrying a variant) and bottom graph shows distribution of statistic S_2 (genes containing a variant in two or more unsolved cases) in 10,000 random tests based on neighbourhood size. S_2 is significant.

neighbour of *RBPJ*. However, while this gene survives filtering at level 5, the same is not true at level 4 because very rare or novel variants are observed too frequently in the non-AOS control exomes.) These genes could be investigated further, but again due to the ambiguous functional relevance of neighbours (especially indirect neighbours) in the co-expression network this would not be prioritised ahead of following up the PINAmin2 and Multinet results described earlier.

On a technical note, it should be pointed out that the permutation tests include a slight bias. This is because each neighbourhood is tested against 10,000 randomly selected gene sets of the same size. However, we know that the seed gene in the original neighbourhood, and possibly one or more neighbours (other known AOS genes; see Table 7.3), cannot contain a variant in an unsolved AOS exome. This bias could be corrected by factoring in the number of known AOS genes when calculating the size of the random gene sets to sample. However, since the bias acts conservatively (the expected p-values would be lower for the corrected test) the results presented here are valid.

This analysis has demonstrated that a network-based candidate-gene approach, such as a simple neighbourhood search around known disease-causing genes, can identify

promising genes for follow-up study. In particular, we identified several interesting genes due to their network proximity to *ARHGAP31*. The missense variants identified in the Rho GTPase signalling genes *ITSN1*, *KALRN*, *SOS1* and *TIAM1*, along with the novel variants in *MAP3K11* (one of which being a nonsense variant), are good candidates to be studied further.

7.3.2 Post-Filtering Variants in Single Genes

The subsequent results sections (7.3.3 to 7.3.6) will present the results of BioGranat-IG analyses, which take a hypothesis-free approach to overcoming genetic heterogeneity. In these analyses the aim is to find small connected subnetworks that harbour post-filtering variants for as many AOS exomes as possible, with no preference given to any genes in the network based on prior knowledge. It is important to view any subnetworks found with a preliminary awareness of why these genes might appear in the results. Therefore it is sensible to look here at the individual genes that harbour post-filtering variants in the highest number of AOS cases. Note that we are essentially performing simple intersection filtering; it may be informative in its own right but is presented here to help understand the BioGranat-IG results that will follow.

For each of filtering levels 1-4, Table 7.5 lists the genes in which the most AOS exomes carry a post-filtering sequence variant. It is not practical to look at every gene in this table individually, but it may be instructive to consider the top few genes.

CNTNAP3B has post-filtering variants in seven exomes at filtering level 1. This number falls to four at filtering level 2, implying that the variants in the other three cases did not match the expected mode of inheritance. Since there are also variants in four exomes at levels 3 and 4, all four of these exomes must be unsolved cases. In fact, *CNTNAP3B* contains heterozygous missense SNVs in exomes S0069 and S0307, two missense SNVs in S0308 that might be compound heterozygous and a novel heterozygous splice-site variant in S0338. The role of *CNTNAP3B* is relatively poorly understood; it is a paralog of *CNTNAP1* which encodes the contactin-associated protein CNTP1 and is associated with several diseases of the nervous system (www.genecards.org, Stelzer et al. 2011). Of the 336 non-AOS control exomes, 19 carry a variant in *CNTNAP3B* (using the criteria for filtering level 1). This is just below the exclusion threshold of 20, suggesting that *CNTNAP3B* may be relatively tolerant to functional variation. These observations suggest that *CNTNAP3B* may not be a good candidate for further study. In terms of its effect on BioGranat-IG results, the only one of the five hub-free networks that contains *CNTNAP3B* is COXPRES30_d50.

CEL has variants in six exomes at filtering level 1, dropping to five at filtering level 2 (implying the variant in the other exome did not match the expected mode of

Table 7.5 – Genes with variants in the highest number of AOS exomes, by filtering level
* = note that filtering level 3 excludes all but the true causal variants for the 6 solved AOS cases.

	Filtering level 1 (from 19 exomes)	**Filtering level 2** (from 19 exomes)	**Filtering level 3** (from 13+6 exomes*)	**Filtering level 4** (from 13 exomes)
7 exomes	*CNTNAP3B*			
6 exomes	*CEL*			
5 exomes	*AGAP6*, *NINL*, *RGPD3*, *USP32*, *ZNF492*	*CEL*		
4 exomes	*ABCA7*, *ARHGAP4*, *ATN1*, *CCDC40*, *CDT1*, *CPVL*, *ITPR3*, *LPA*, *PCDH12*, *PHKB*, *PPM1E*, *PRRC2B*, *RTL1*, *SIPA1L3*, *TPSAB1*, *VCX3B*, *ZDHHC13*, *ZNF493*	*CNTNAP3B*, *COL6A3*, *ITPR3*, *LPA*, *NINL*, *RGPD3*	*CNTNAP3B*, *NINL*	*CNTNAP3B*, *NINL*
3 exomes	91 genes (excluding *LRP2*)	44 genes (including *LRP2*)	*BHLHE22*, *BMP5*, *CILP*, *COL6A3*, *GPR124*, *GPR125*, *LPA*, *LRP2*, *NHSL1*, *PHKB*, *PRB1*, *PTPRG*, *RGPD3*, *ZDHHC13*, *ZNF335*	*BHLHE22*, *BMP5*, *CILP*, *COL6A3*, *GPR124*, *GPR125*, *LPA*, *LRP2*, *NHSL1*, *PHKB*, *PRB1*, *PTPRG*, *RGPD3*, *ZDHHC13*, *ZNF335*

inheritance). Between filtering levels 2 and 3 the number of exomes carrying a variant falls from five to two (not shown in table). This means that only two of the exomes are unsolved cases, the other three having known AOS-causing variants in other genes. Since we assume AOS is monogenic, we do not expect any additional variants in the solved cases to be linked to AOS and *CEL* will not be considered further. Note that *CEL* is present in PINA_d50 and Multinet_d50 so may influence BioGranat-IG results in these networks.

NINL has variants in five exomes at filtering level 1, dropping to four at level 2 (implying one exome did not have a variant that matched the expected mode of inheritance). Since there are also variants in four exomes at filtering levels 3 and 4 these must all be unsolved cases. In fact *NINL* contains (the same) novel heterozygous splice-site variant in exomes S0307 and S0338, and novel heterozygous missense SNVs in S0308 and S0339. *NINL* encodes ninein-like protein, which is involved in microtubule organisation in interphase cells. Of the 336 non-AOS control exomes, eight carry a variant in *NINL* (using

the criteria for filtering level 1). This is slightly higher than average (across all 16,282 genes with a post-filtering variant in at least one control the mean is 5.80), but not exceptionally high: 2,382 genes harbour post-filtering variants in a higher number of non-AOS control exomes. Note that *NINL* is present in all hub-free networks except PINAmin2_d50 and so could influence BioGranat-IG results.

Other genes will be examined as they occur in BioGranat-IG results.

7.3.3 BioGranat-IG Results: Summary

In total, 80 different BioGranat-IG searches were performed using each combination of five different networks (PINA_d50, PINAmin2_d50, CPDBconf95_d50, COXPRES30_d50 and Multinet_d50), four different search methods (exact triplet and quadruplet searches, heuristic search limited to ten genes, unlimited heuristic search) and four different variant filtering levels (1-4). Table 7.6 summarises the findings for each search, giving the number of genes in an optimal subnetwork and the number of AOS exomes in which an optimal subnetwork harbours a variant. Note that optimal subnetworks are not necessarily unique because several equivalently good subnetworks might be found in a given search (optimality implies that no subnetworks were found to harbour variants in a greater number of AOS exomes and no smaller subnetworks were found to harbour variants in the same number of AOS exomes). For the triplet and quadruplet searches, near-optimal subnetworks are also considered.

Since it is not practical to examine in detail every subnetwork found, the following sections will take a logical approach in order to highlight subnetworks of greatest potential interest. Most attention is given to the searches in the PINA_d50 network since this is the network of physical interactions with widest genomic coverage (10,375 genes; COXPRES30_d50 includes a greater number of genes but interactions in this network are less easily interpretable). Subsequently, findings will be compared to the PINAmin2_d50 and CPDBconf95_d50 networks, which are both smaller PINs than PINA_d50 but whose interactions are of higher confidence on average. Finally KGGSeq-prioritisation, as described in chapter 6, will be used to identify any additional subnetworks from the remaining networks that warrant further attention.

7.3.4 BioGranat-IG Results: PINA_d50 Network

7.3.4.1 Filtering Level 4

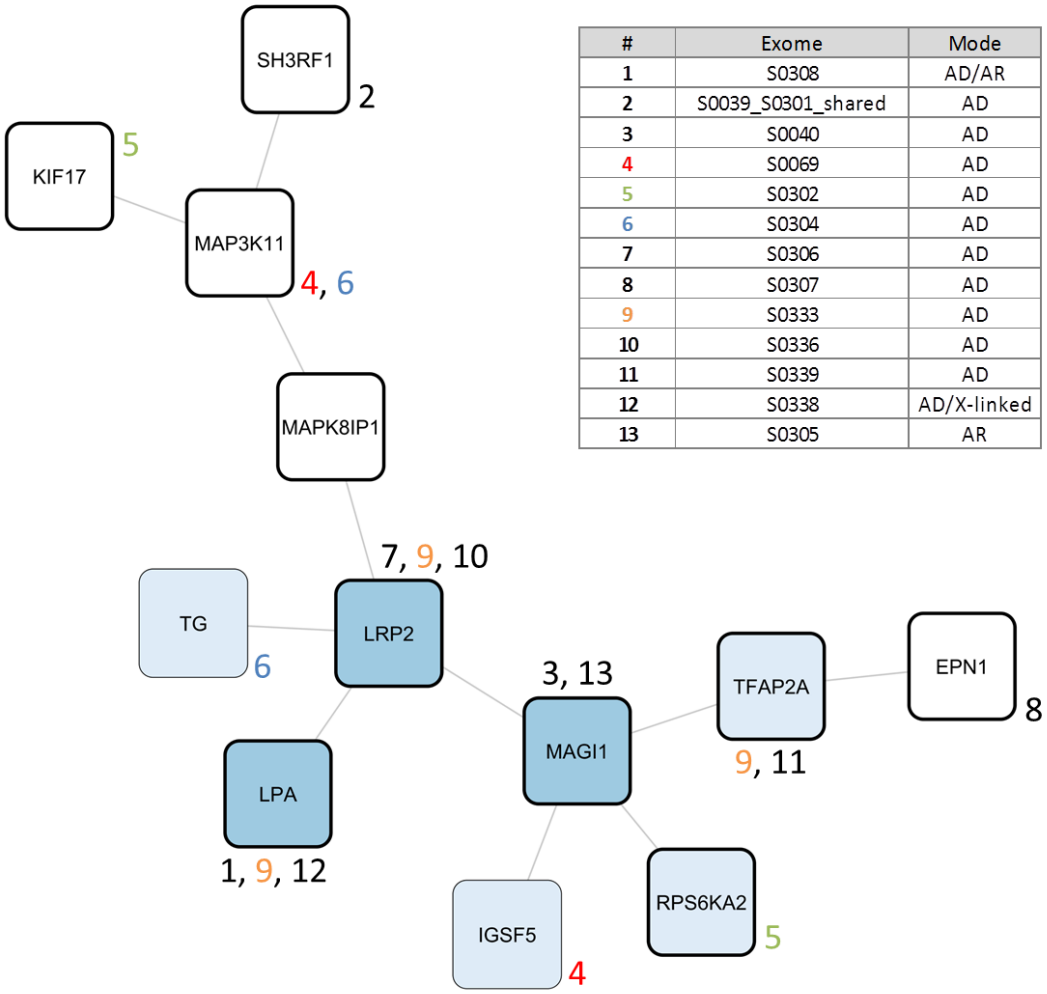
Filtering level 4 includes novel and very rare non-synonymous variants that match the expected mode of inheritance in the 13 unsolved AOS cases. It is interesting to start with the results of this analysis, which does not make use of the exome data from solved AOS

Table 7.6 – Summary of optimal subnetworks for AOS found by BioGranat-IG
Results take the form: *number of AOS cases covered* (*number of genes*). Note this table gives the properties of optimal subnetworks found by each search but optimal subnetworks are not necessarily unique.

Network	Triplet search	Quadruplet search	Heuristic (limit 10)	Heuristic (unlimited)
Filtering level 1 *(from 19 exomes)*				
PINA_d50	8 (3)	10 (4)	17 (10)	19 (12)
PINAmin2_d50	6 (3)	7 (4)	12 (9)	19 (18)
CPDBcon95_d50	7 (3)	9 (4)	15 (10)	19 (16)
COXPRES30_d50	9 (3)	10 (4)	18 (10)	19 (12)
Multinet_d50	10 (3)	11 (4)	17 (10)	19 (13)
Filtering level 2 *(from 19 exomes)*				
PINA_d50	9 (3)	10 (4)	17 (10)	19 (12)
PINAmin2_d50	5 (3)	7 (4)	14 (10)	19 (15)
CPDBcon95_d50	7 (3)	9 (4)	14 (10)	19 (17)
COXPRES30_d50	7 (3)	9 (4)	15 (10)	19 (15)
Multinet_d50	9 (3)	10 (4)	17 (10)	19 (13)
Filtering level 3 *(from 19 exomes)*				
PINA_d50	7 (3)	8 (4)	13 (9)	14 (12)
PINAmin2_d50	5 (3)	6 (4)	11 (10)	17 (20)
CPDBcon95_d50	5 (3)	6 (4)	11 (10)	17 (19)
COXPRES30_d50	6 (3)	7 (4)	11 (10)	19 (26)
Multinet_d50	7 (3)	7 (3)	12 (10)	16 (14)
Filtering level 4 *(from 13 exomes)*				
PINA_d50	7 (3)	8 (4)	13 (9)	13 (9)
PINAmin2_d50	5 (3)	6 (4)	11 (10)	13 (13)
CPDBcon95_d50	5 (3)	6 (4)	11 (10)	13 (12)
COXPRES30_d50	6 (3)	7 (4)	12 (9)	13 (11)
Multinet_d50	7 (3)	7 (3)	13 (10)	13 (10)

cases, because BioGranat-IG has the capacity to suggest novel disease mechanisms. It is already established that variants in different functional pathways can cause AOS (for example, *ARHGAP31* and *DOCK6* cause AOS through a consistent molecular mechanism (Shaheen et al. 2011) with *NOTCH1* and *RBPJ* through another (Stittrich et al. 2014)); while the simple neighbourhood search has the potential to find new genes causing AOS through an existing mechanism, BioGranat-IG can propose causal mechanisms that are not necessarily related to the known AOS pathways.

At filtering level 4, the optimal triplet found by BioGranat-IG in PINA_d50 contains a variant in seven AOS exomes (depicted in Figure 7.7). *LPA* contains a variant in exomes S0308, S0333, and S0338; *LRP2* also contains a variant in exome S0333, along with S0306 and S0336, and *MAGI1* contains a variant in S0040 and S0305. The subnetwork has a



#	Exome	Mode
1	S0308	AD/AR
2	S0039_S0301_shared	AD
3	S0040	AD
4	S0069	AD
5	S0302	AD
6	S0304	AD
7	S0306	AD
8	S0307	AD
9	S0333	AD
10	S0336	AD
11	S0339	AD
12	S0338	AD/X-linked
13	S0305	AR

Figure 7.7 – Summary of BioGranat-IG results for PINA_d50 at AOS filtering level 4
Dark blue: optimal triplet; adding any one light blue gives optimal quadruplet; bold border: optimal heuristic (*TG* and *IGSF5* not included because *MAP3K11* already has a variant in the same exomes; *KIF17* and *RPS6KA2* are alternative choices in the optimal heuristic subnetwork because both cover exome S0302). The number(s) next to each node refer to the exomes in which they contain variants; coloured numbers indicate exomes with multiple variants in the subnetwork.

significant KGGSeq-prioritisation score when compared against 100,000 sets of randomly selected post-filtering variants in the same exomes (probability of containing a disease-causing variant = 0.8361, $p = 0.0221$). This score is driven by the high probability (0.6950) of causing some monogenic disease assigned by KGGSeq to a novel missense SNV found in *LRP2* in exome S0333.

However, there is also evidence suggesting that this triplet might not represent a subnetwork involved in AOS. Firstly, exome S0333 has a variant in both of the genes *LPA* and *LRP2*. Our analysis assumes that AOS is monogenic and therefore we would not expect our subnetwork to contain variants in different genes in the same exome. (Note this does not

definitively rule out this subnetwork: if these genes form a functional pathway underlying AOS then two less severe mutations in different genes could feasibly result in a similar phenotypic outcome to one more severe mutation in a single gene; alternatively, one of the two variants in S0333 could have no functional effect and be present by chance.) Secondly, the gene *LRP2* appears relatively tolerant to functional variation in the 336 non-AOS control exomes. While only 17 of the 336 exomes carry post-filtering heterozygous variants (therefore not meeting the exclusion threshold of 20 for these exomes at filtering levels 2-4), this number rises to 49 when homozygous or compound heterozygous variants are also included (and hence *LRP2* is actually excluded at filtering level 1).

There are four alternative optimal quadruplets. Each can be formed by extending the optimal triplet by one gene to find a variant in one additional exome (see Figure 7.7). The genes are *IGSF5* (variant in S0069), *RPS6KA2* (variant in S0302), *TG* (variant in S0304) and *TFAP2A* (variant in S0339, as well as in S0333 which was already covered by the optimal triplet). Individually, each optimal quadruplet has a significant KGGSeq-prioritisation score, but the subnetwork found by merging all four quadruplets does not ($p = 0.0990$).

The heuristic searches find two overlapping optimal subnetworks of nine genes in which all 13 unsolved AOS exomes have a variant. Since the optimal subnetwork has fewer than ten genes the unlimited heuristic search gives the same results as the heuristic search limited at size ten. Merging the two subnetworks gives a network region of ten genes, which include the three genes from the optimal triplet and two of the additional optimal quadruplet genes (see Figure 7.7). The KGGSeq-prioritisation score for the merged region is significant (probability of containing a disease-causing variant = 0.9857, $p = 0.0050$).

The triplet, quadruplet and heuristic searches have identified a network region of 12 genes in total because there is a concentration of post-filtering variants carried by the unsolved AOS exomes in this part of the network. It is therefore natural to examine whether these genes are known to play a role in some common functional process. Three significant GO biological process annotations were identified. Four genes (*MAP3K11*, *MAPK8IP1*, *RPS6KA2* and *SH3RF1*) are involved in the "stress-activated MAPK cascade" (p-value after adjustment for multiple testing: $adjP = 0.0026$). This is of potential interest given the critical importance of spatially- and temporally-precise cell signalling to normal development (Southgate et al. 2011): the known AOS genes *ARHGAP31* and *DOCK6* regulate the Cdc42 and Rac1 signalling processes (Shaheen et al. 2011; Southgate et al. 2011) while *NOTCH1* and *RBPJ* are involved in notch-mediated signalling (Hassed et al. 2012; Stittrich et al. 2014). In particular, cross-talk between MAPK and notch signalling pathways have been demonstrated, although the relationship is not fully understood (Kondoh et al. 2007; Yamashita et al. 2013). Four genes (*MAP3K11*, *RPS6KA2*, *SH3RF1* and *TFAP2A*) are

involved in the related process "positive regulation of apoptotic process" (*adjP* = 0.0105). Six genes (*EPN1*, *KIF17*, *LPA*, *LRP2*, *MAPK8IP1* and *TG*) are involved in "vesicle-mediated transport" (*adjP* = 0.0105).

The 12 genes identified almost certainly do not represent a previously unknown full explanatory functional pathway underlying AOS. The main argument for this is network coverage: of the 1,983 genes in which post-filtering variants were identified across all 13 unsolved AOS exomes only 1,087 (54.8%) map to genes in PINA_d50. Network coverage is limited both by having removed hub genes and by the fact that current knowledge of the interactome is far from complete (Yu et al. 2011). In addition, we should consider the secondary problem of whole exome sequencing coverage: Table 7.1 lists exome capture scores as low as 71.0% at 20× coverage, meaning that for some of the exomes the true AOS-causing variant may not even have been sequenced to a sufficient depth to have been correctly identified (we know for example that this was the case for one of the *DOCK6* mutations that cause AOS in exome S0334). However, BioGranat-IG is designed to carry out a specific search and is unaware of these wider considerations; its unlimited heuristic search in particular will continue to add additional genes to promising subnetworks until it has found variants for all of the exomes, meaning that sequencing coverage is more problematic for BioGranat-IG than it might be for simple intersection filtering. (This leaves aside the possibility that since the minimum distance and multi-minimum distance searches are not exact they could fail to identify a smaller subnetwork in which all exomes carry a variant.)

There are also features specific to this set of 12 genes that should be noted. Four of the 13 unsolved AOS cases carry more than one variant in the subnetwork, and as described previously only one per exome would be expected to be causal.

One gene (*MAPK8IP1*) contains no post-filtering variant in any of the exomes. On one hand this means there is no direct evidence that would link this gene to AOS, but on the other hand this could be explained by sampling effects (variants in this gene can cause AOS but not for any of the affected individuals we have sequenced) or because the gene is too critical to the pathway to tolerate mutations of severe effect (or not critical enough to cause the AOS phenotype, for example due to the existence of paralogous genes).

The 12 genes are connected in the network by 11 edges, the minimum number required to connect 12 nodes. There is some evidence to suggest that densely-connected network regions are more likely to be involved in disease (Garcia-Alonso et al. 2012).

None of the 12 genes in this network region interact directly or indirectly with a known AOS gene in PINA_d50. (However, in section 7.3.1 we saw that *MAGI1*, *MAP3K11*

and *SH3RF1* were part of the neighbourhood of *ARHGAP31* in the Multinet network that was significantly enriched for post-filtering variants in the simple neighbourhood search.)

Although the full network region is unlikely to be an AOS disease pathway, it is still possible that some part of it plays a role in AOS. Notwithstanding the potentially interesting finding of genes involved in MAPK signalling at the periphery of the region, the most promising genes are perhaps the optimal triplet genes (and to a lesser extent the optimal quadruplet genes) that form the core of the region, due to the concentration of variants across the AOS exomes.

To investigate other small regions containing a high concentration of post-filtering variants the triplet and quadruplet searches were repeated to search for near-optimal results. Since the optimal triplet contained a variant for seven AOS cases, near-optimal triplets must cover six AOS cases and near-optimal quadruplets must cover seven. The near-optimal triplet search identified triplets in two distinct regions. Firstly, eight triplets, each containing at least two of *LPA*, *LRP2* and *MAGI1* give a region of nine genes when merged, covering 11 AOS cases. These genes largely overlap with the network region already discussed and will not be discussed further. Secondly, seven triplets covering six AOS cases each can be merged to give a region of ten genes covering ten AOS cases. However, these results are driven by the gene *NINL* which contains a variant in four AOS cases (as discussed in section 7.3.2 above). Therefore whether this region could plausibly be an AOS disease pathway rests mainly on the question of whether *NINL* is an AOS gene.

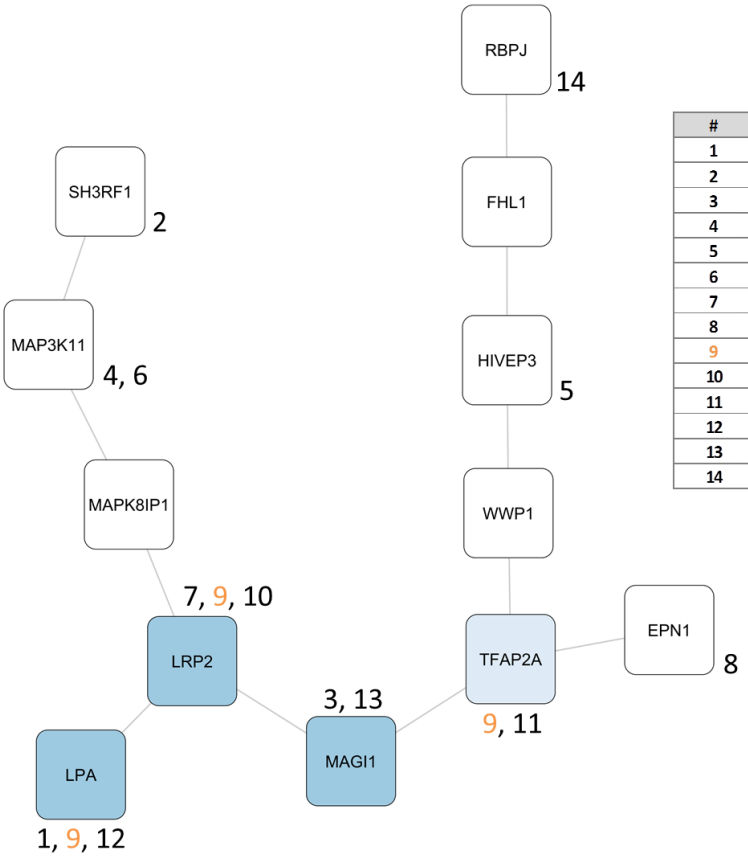
The near-optimal quadruplet search identified 57 quadruplets which form a merged region of 47 genes (including *LPA*, *LRP2* and *MAGI1*, as well as *NINL*) covering all 13 AOS cases. It is difficult to draw conclusions about the AOS disease process by studying a network region of this size.

7.3.4.2 Filtering Level 3

Filtering level 3 is the same as filtering level 4 except that the six solved AOS cases are also included, but with only the true causal variants being included for each.

The optimal triplets and optimal quadruplets found in PINA_d50 at filtering level 3 were identical to those found at level 4. This implies that there are not sufficient post-filtering variants in the vicinity of any of the known AOS genes to form highly-mutated triplets or quadruplets. Even though we have implicitly weighted the search towards the known AOS genes, the small region around *LPA*, *LRP2* and *MAGI1* still has the strongest evidence for AOS involvement, in terms of variants in AOS exomes.

This is true to the extent that even the near-optimal triplets and quadruplets found at filtering level 3 are the same as those found at level 4.



#	Exome	Mode
1	S0308	AD/AR
2	S0039_S0301_shared	AD
3	S0040	AD
4	S0069	AD
5	S0302	AD
6	S0304	AD
7	S0306	AD
8	S0307	AD
9	S0333	AD
10	S0336	AD
11	S0339	AD
12	S0338	AD/X-linked
13	S0305	AR
14	S0337	RBPJ only

Figure 7.8 – Optimal subnetwork found in PINA_d50 at AOS filtering level 3 using unlimited heuristic searches
Dark blue: overlap with optimal triplet; light blue: also in one of the optimal quadruplets. The number(s) next to each node refer to the exomes in which they contain variants; coloured numbers indicate exomes with multiple variants in the subnetwork.

For the minimum distance and multi-minimum distance searches, when these were limited at size ten the same two sets of nine genes were found as for filtering level 4. However, when no limit was placed on subnetwork size BioGranat-IG identified an optimal subnetwork of 12 genes in which 14 of the 19 AOS exomes carried a variant (depicted in Figure 7.8). This region contains variants for all 13 unsolved cases, as well as the true causal variant for exome S0337 in the gene *RBPJ*. There is substantial overlap (eight genes) between this network region and the one previously identified at filtering level 4 and presented in Figure 7.7. Both regions contain the core triplet of *LPA*, *LRP2* and *MAGI1*, plus a path of interactions connecting *LRP2* up to *SH3RF1* via *MAP3K11* and one leading out to *EPN1* from *MAGI1*. However, at filtering level 3 a path of length four is added from the gene *TFAP2A* to incorporate *RBPJ*. This path includes two connecting genes (*FHL1* and *WWP1*) which do not harbour a post-filtering variant in any of the AOS exomes. Based on the fact that so few variants are observed in this path it seems unlikely that there is a genuine AOS-relevant functional link between *RBPJ* and the previously identified genes. Rather, this

path has been identified because BioGranat-IG's objective is simply to connect together variants in as many AOS exomes as it can.

However, existing functional annotation suggests that *RBPJ* may be more closely linked to the previously identified genes than inspection of this network region might suggest. An enrichment test based on the optimal subnetwork genes found by the unlimited heuristic search (the genes in Figure 7.8) identified two significant GO biological process annotations. Three of the four genes found earlier to play a role in the "stress-activated MAPK cascade" remain in this region (*MAP3K11*, *MAPK8IP1* and *SH3RF1*; *adjP* = 0.0194). In addition three genes are involved in "regulation of ERBB signalling pathway": *EPN1*, *RBPJ* and *TFAP2A* (*adjP* = 0.0066).

Also of note is the fact that the KGGSeq-prioritisation score is significant for this region (probability of containing a disease-causing variant = 0.9848, $p = 0.0047$). Of course we know this region does contain a disease-causing variant in *RBPJ* for exome S0337. KGGSeq ascribes this variant a probability of 0.1723 of causing a monogenic disease, which is relatively high (92.44% of the variants across all exomes at filtering level 3 have a lower probability). However, the significant score is driven by a probability of 0.8604 assigned to a novel missesnse SNV found in *MAP3K11* in exome S0304 and a probability of 0.6950 assigned to a novel missense SNV found in *LRP2* in exome S0333.

Finally, it is worth noting why BioGranat-IG did not find a region which included the genes *ARHGAP31*, *DOCK6* and *NOTCH1*, since these genes contain post-filtering variants in the five additional exomes not covered by the region depicted in Figure 7.8. This is because, although all three of these genes are present in the full PINA network, none remain in PINA_d50 after hub removal. *NOTCH1* is removed because it is itself a hub gene (65 interaction partners in PINA), while *ARHGAP31* and *DOCK6* are only connected to the rest of the network via hub genes so are lost when these hubs are removed. Table 7.7 summarises the presence of the known AOS genes in each of the five hub-free networks.

7.3.4.3 Filtering Level 2

At filtering level 2, all 19 AOS exomes are subject to the same filtering steps, giving all novel and very rare non-synonymous variants that match the expected mode of inheritance for each AOS case. The only difference to filtering level 3 is that the solved cases are now represented by a full list of filtered variants and not just the variants we know to cause AOS. This allows us to simulate a scenario in which the true causal AOS variants have not yet been identified for the six solved cases, and to ask whether BioGranat-IG can pick out these true causal variants. We can also examine how the "noise" introduced by the non-causal variants in the solved cases affects the subnetworks that BioGranat-IG finds.

Table 7.7 – Presence of known AOS genes in hub-free networks

Network	Known AOS gene	In network?	Notes
PINA_d50	*ARHGAP31*	✗	Only connected to hub genes in full network
	DOCK6	✗	Only connected to hub genes in full network
	EOGT	✗	Not in full network
	NOTCH1	✗	Hub gene in full network
	RBPJ	✓	
PINAmin2_d50	*ARHGAP31*	✓	
	DOCK6	✗	Only connected to hub genes in full network
	EOGT	✗	Not in full network
	NOTCH1	✓	*RBPJ* is a direct neighbour
	RBPJ	✓	*NOTCH1* is a direct neighbour
CPDBconf95_d50	*ARHGAP31*	✓	
	DOCK6	✗	Not in full network
	EOGT	✗	Not in full network
	NOTCH1	✓	*RBPJ* is a direct neighbour
	RBPJ	✓	*NOTCH1* is a direct neighbour
COXPRES30_d50	*ARHGAP31*	✓	
	DOCK6	✓	
	EOGT	✗	Not in full network
	NOTCH1	✓	
	RBPJ	✓	
Multinet_d50	*ARHGAP31*	✗	Only connected to hub genes in full network
	DOCK6	✗	Only connected to hub genes in full network
	EOGT	✗	Not in full network
	NOTCH1	✓	*RBPJ* is a direct neighbour
	RBPJ	✓	*NOTCH1* is a direct neighbour

Of course, there are two reasons we know in advance that BioGranat-IG will not be able to pick out all of the known AOS genes. Firstly, since one of the true causal mutations in *DOCK6* for exome S0334 was not captured by the whole exome sequencing process this gene is not represented in the gene lists at filtering level 2. Secondly, PINA_d50 does not cover the genes *ARHGAP31*, *DOCK6* or *NOTCH1*.

The optimal triplet found by BioGranat-IG in PINA_d50 is the same as that found for filtering levels 3 and 4. It consists of the genes *LPA*, *LRP2* and *MAGI1*. However, when variants are filtered at level 2 this triplet now covers two additional AOS exomes (nine in total). Exome S0038 has a variant in *LPA*, although we know in reality that a mutation in *ARHGAP31* causes AOS for this individual; likewise exome S0335 carries a variant in *MAGI1* but in this case AOS is caused by *NOTCH1*.

Similarly, the results of the optimal quadruplet search at filtering level 2 are broadly comparable with those found at levels 3 and 4. It is now possible to cover ten of the 19 AOS exomes using a four-gene subnetwork, and eight such optimal quadruplets were identified. However, in the region of 11 genes formed by merging those quadruplets, seven are the same genes identified by quadruplet search at filtering levels 3 and 4, while the other four are only present due to variants in the six solved cases that we know do not cause AOS. In fact, two of the new genes (*DAB1* and *PIP5K1C*) are identified because they harbour a variant in exome S0337, for which the true AOS-causing variant can be found in *RBPJ* elsewhere in the network. This clearly demonstrates a limitation of BioGranat-IG in dealing with "noisy" exome sequencing data: these genes are picked up instead of *RBPJ* because they are closer (in network terms) to the concentration of variants in the other exomes around the genes *LPA*, *LRP2* and *MAGI1*.

When the heuristic searches were performed with no size limit they were are able to find a subnetwork of 12 genes in which all 19 of the AOS exomes contain a variant at filtering level 2. The subnetwork does not add any new insight: it is based around *LPA*, *LRP2* and *MAGI1*; it does not have a significant KGGSeq-prioritisation score ($p = 0.1269$); it is not densely connected, having the minimum number of edges (11) required to be connected, and all six of the solved AOS cases are represented by variants that we know are not causal (including S0337, the only one where the true causal gene is in PINA_d50 but which is covered in this subnetwork by *DAB1*).

Likewise, the heuristic searches limited to subnetworks of size ten add little. Three optimal subnetworks of ten genes were found, each covering 17 of the 19 AOS exomes. Common to all three are *DAB1*, *LRP2*, *MAGI1* and *MYO6*, and two of the subnetworks also cover *NINL* (which contains a variant for four of the unsolved exomes as discussed in section 7.3.2). The known AOS gene *RBPJ* is not covered. As before, the differences between these results and the heuristic search results at filtering levels 3 and 4 are due to variants in the solved cases which we know are not actually causal (under the assumption that only one variant per exome causes AOS).

7.3.4.4 Filtering Level 1

By examining the BioGranat-IG results at filtering level 1 we can see the effects of relaxing the filtering criterion that requires variants to match the expected mode of inheritance in each AOS case. One initial observation is that variants in *LRP2* are excluded at filtering level 1 (as discussed when the filtering level 4 results were presented in section 7.3.4.1 above, *LRP2* contains post-filtering variants in more than 20 of the 336 non-AOS control exomes when mode of inheritance is disregarded). This means the results are not

expected to match closely what we saw at filtering levels 2-4, where *LRP2* was central to most of the optimal subnetworks.

Optimal triplets in PINA_d50 harbour variants in eight AOS cases at filtering level 1. There are 13 such optimal triplets, which form two distinct merged regions in the network (see Figure 7.9). Ten include the gene *NINL*, which has been previously discussed and contains variants in five AOS exomes at filtering level 1 (one of which is the solved case S0332). In the merged region, 13 genes cover 15 of the AOS cases in total. There are six surplus variants (that is, the region contains 21 variants in total for these 15 cases) and the region is not densely connected (it has the minimum number of edges required for connectivity), each of which arguably points away from a disease role. Neither the merged region nor any of the individual triplets has a significant KGGSeq-prioritisation score. This region is almost certainly not of relevance to AOS unless the individual gene *NINL* is. The second merged region is formed by the remaining three triplets, all of which include the genes *CEL* and *LTF*. As described in section 7.3.2, *CEL* contains variants in six exomes at filtering level 1, and again whether this merged region is of relevance to AOS is highly dependent on this gene.

The optimal quadruplet search results are consistent with the optimal triplets. Again, the five optimal quadruplets (which cover 10 AOS cases each) form two distinct merged regions in the network, one based around *NINL* and one based around *CEL* (see Figure 7.9), and do not warrant further discussion.

Interestingly, the optimal subnetworks identified by the heuristic searches (both when limited to subnetworks of size ten and when executed with no limit) contain the genes *LPA*, *LRP2* and *MAGI1*, despite the fact that *LRP2* variants are excluded at filtering level 1. Further, there is no overlap between the optimal triplets and quadruplets and the results from the heuristic searches (see Figure 7.10). However, other than the three genes *LPA*, *LRP2* and *MAGI1*, there are no other genes in common with any of the regions found at filtering level 4 (that is, with any of the genes in Figure 7.7). The presence of *LPA*, *LRP2* and *MAGI1* is most likely explained by the fact that *LRP2*, with 41 interaction partners, is relatively highly connected in the PINA_d50 network (with degree 49 in the full PINA network, *LRP2* is just short of the threshold for removal as a hub gene). Nodes of high degree are more likely than lower-degree nodes to be identified in BioGranat-IG results as connecting genes because of the higher probability that two or more of the neighbouring genes will contain post-filtering variants. *LPA* contains post-filtering variants in four AOS cases, and *MAGI1* in three, although for each gene one is a solved case, having a causal variant in a known AOS gene. There is little other evidence that the regions identified by the heuristic searches are involved in AOS: the KGGSeq-prioritisation scores are far from significant for both the size-

#	Exome
1	***S0038***
2	S0039_ S0301_ shared
3	S0040
4	S0069
5	S0302
6	S0304
7	S0305
8	S0307
9	S0308
10	***S0311***
11	***S0332***
12	S0333
13	***S0334***
14	***S0335***
15	S0336
16	***S0337***
17	S0338
18	S0339

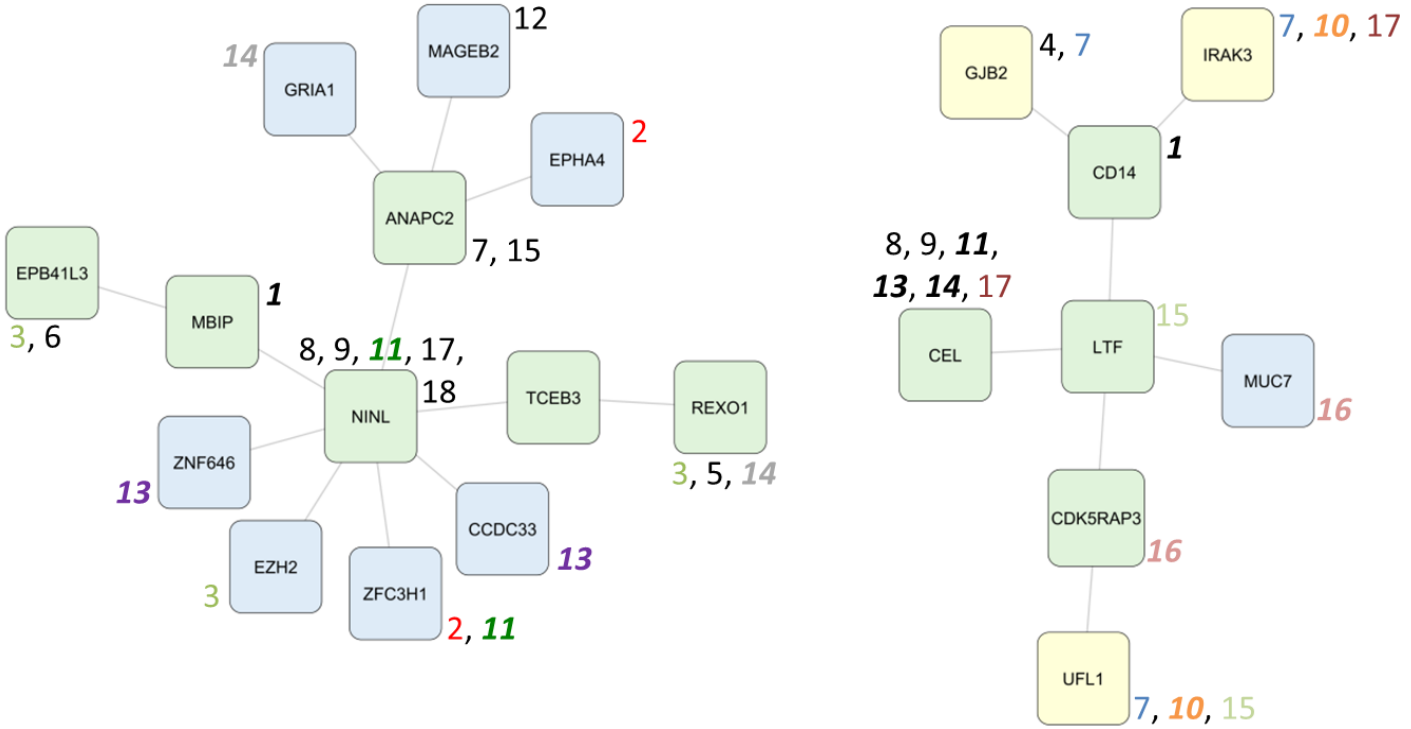
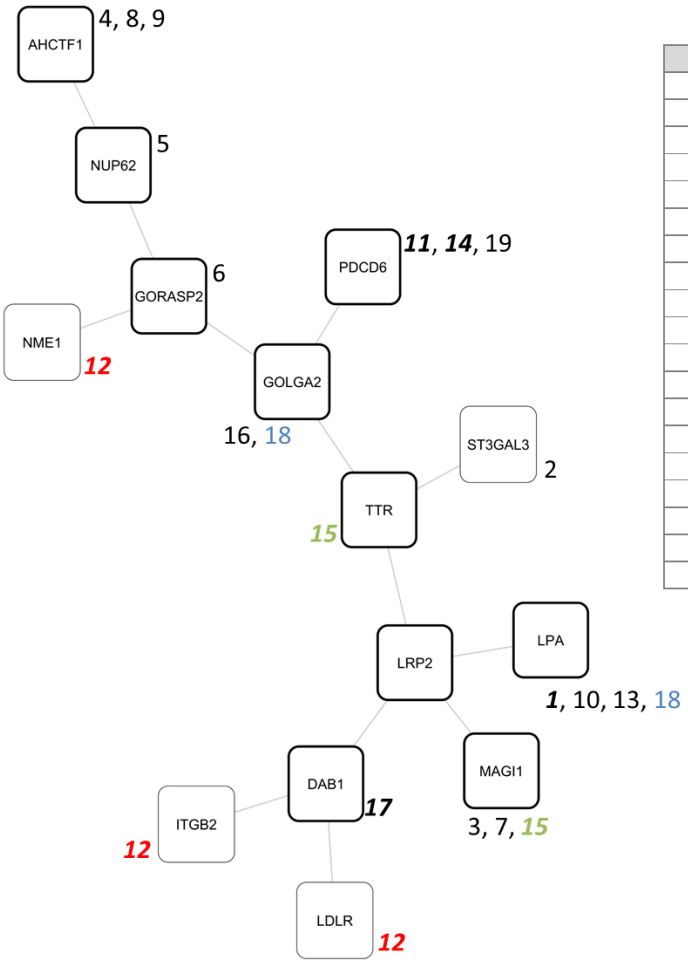


Figure 7.9 – Optimal subnetworks found in PINA_d50 at AOS filtering level 1 using triplet and quadruplet searches
Merged regions shown. Optimal triplets shown in blue; quadruplets in yellow; overlap in green. The number(s) next to each node refer to the exomes in which they contain variants; bold italic indicates a solved AOS case with causal mutations found elsewhere; coloured numbers indicate exomes with multiple variants in the same merged region.



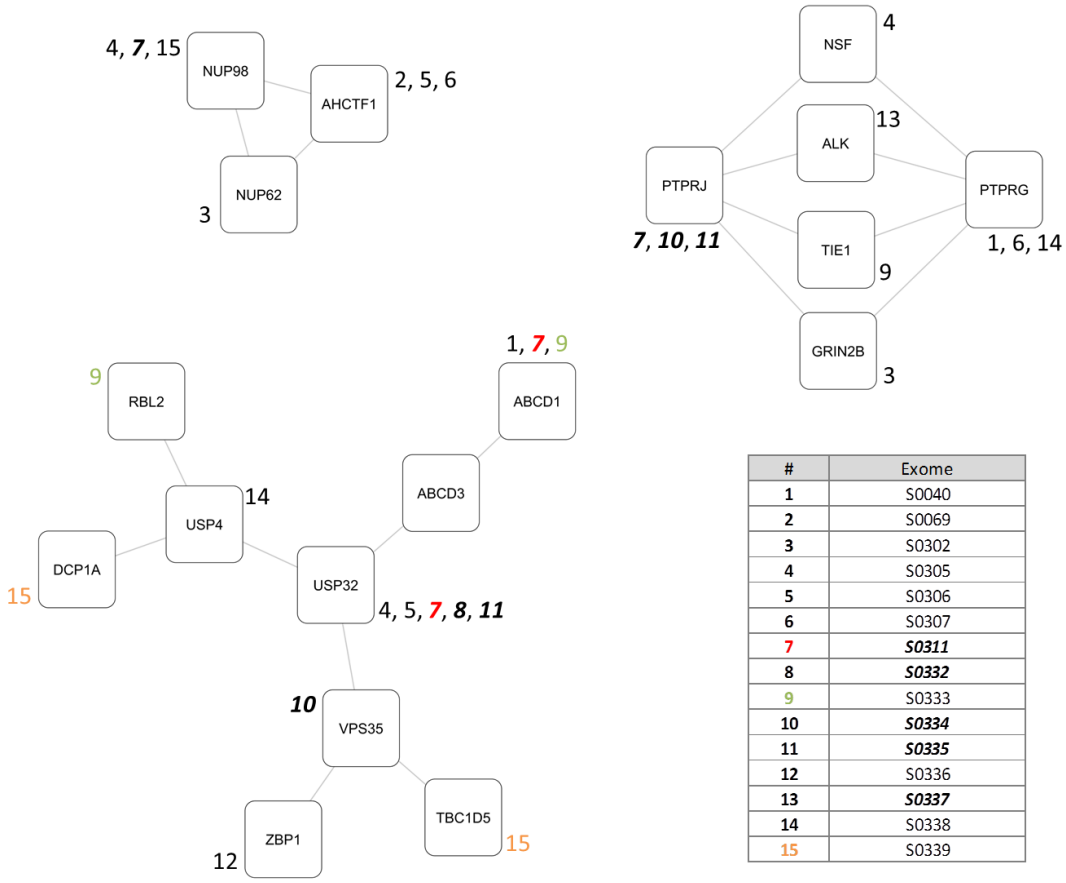
#	Exome
1	***S0038***
2	S0039_S0301_shared
3	S0040
4	S0069
5	S0302
6	S0304
7	S0305
8	S0306
9	S0307
10	S0308
11	***S0311***
12	***S0332***
13	S0333
14	***S0334***
15	***S0335***
16	S0336
17	***S0337***
18	S0338
19	S0339

Figure 7.10 – Optimal subnetworks found in PINA_d50 at AOS filtering level 1 using heuristic searches
All genes: merged region found by unlimited heuristic searches; bold border: optimal subnetwork when searches limited to subnetworks of ten genes or fewer. The number(s) next to each node refer to the exomes in which they contain variants; bold italic indicates a solved AOS case with causal mutations found elsewhere; coloured numbers indicate exomes with multiple variants in the region.

limited and unlimited network regions, and the 14 genes that make up the regions are connected by the minimum number of edges possible to ensure connectivity (13).

Looking beyond optimal subnetworks, the near-optimal triplet search results give an indication of which regions of the network are enriched for closely connected variants after filtering at level 1. A near-optimal triplet harbours a variant in seven AOS cases, and such triplets were found in six distinct network regions. One is based around the gene *NINL*, one around *CEL*, and one around the *LPA-LRP2-MAGI1* triplet; these genes have been previously discussed. The remaining three network regions, shown in Figure 7.11, are each of interest for different reasons.

The triplet of genes formed of *AHCTF1*, *NUP62* and *NUP98* is fully connected, suggesting a close functional relationship. An enrichment test found significant GO biological process terms "mRNA transport genes" (all three genes; $adjP = 2.38\times10^{-5}$) and



#	Exome
1	S0040
2	S0069
3	S0302
4	S0305
5	S0306
6	S0307
7	***S0311***
8	***S0332***
9	S0333
10	***S0334***
11	***S0335***
12	S0336
13	***S0337***
14	S0338
15	S0339

Figure 7.11 – Three regions of near-optimal triplets found in PINA_d50 at AOS filtering level 1
The number(s) next to each node refer to the exomes in which they contain variants; bold italic indicates a solved AOS case with causal mutations found elsewhere; coloured numbers indicate exomes with multiple variants in the same region. Not shown: three network regions based around the gene *NINL*, the gene *CEL* and the triplet *LPA-LRP2-MAGI1* respectively.

"nuclear pore organisation" (*AHCTF1* and *NUP98*; adjP = 2.38×10^{-5}). Two of the genes, *AHCTF1* and *NUP62*, were previously seen as part of the optimal subnetworks found by the heuristic searches (see Figure 7.10).

Four triplets form a region of six genes, based around *PTPRJ* and *PTPRG*. This region contains three additional edges above the minimum required for connectivity. Note that in PINA_d50, *PTPRJ* has degree 23 and *PTPRG* has degree 15; there are 11 genes connected to both of them (the region identified here contains four of these, each containing one post-filtering variant). The GO enrichment test revealed several functional roles for the genes involved: *ALK*, *GRIN2B* and *PTPRJ* are involved in "regulation of MAPK cascade" (*adjP* = 0.0204); *PTPRG* and *PTPRJ* in "peptidyl-tyrosine dephosphorylation" (*adjP* = 0.0144); *TIE1* and *NSF* in "plasma membrane fusion" (*adjP* = 0.0015), and *TIE1* and *PTPRJ* in "negative regulation of cell motility" (*adjP* = 0.0204). Note that all of the

variants observed in *PTPRJ* come from the exomes of solved AOS cases, and are therefore not expected to be disease-relevant.

Finally, six near-optimal triplets form a network region centred on the gene *USP32*, which contains five post-filtering variants (although three come from solved AOS cases). This region is of interest because it has the most significant KGGSeq-prioritisation score (probability of containing a disease-causing variant = 1.0000, $p < 10^{-5}$). This is driven by the three variants found in the gene *ABCD1*: two (in exomes S0040 and S0311) are annotated with disease-causing probabilities of >0.99 and the other (in exome S0333) has a probability of 0.7938. However, *ABCD1* is found on the X chromosome, and none of these AOS cases are expected to have X-linked mode of inheritance.

7.3.5 BioGranat-IG Results: Higher-Confidence PINs

Having studied in detail the BioGranat-IG results in the network PINA_d50, we now examine whether these are supported by the results found in PINAmin2_d50 and CPDBconf95_d50. These PINs cover a lower proportion of the genome than PINA_d50 but only include higher-confidence interactions. Regions identified by BioGranat-IG in these networks are less likely to contain false positive protein-protein interactions (PPIs) and therefore more likely to represent genuine functional relationships.

7.3.5.1 Filtering Level 4

The region of 12 genes identified in PINA_d50 by the optimal searches at filtering level 4 (as depicted in Figure 7.7) is not found intact in either of the higher-confidence PINs PINAmin2_d50 or CPDBconf95_d50 (see Figure 7.12). This means that BioGranat-IG will not pick out precisely the set of genes seen previously.

In the PINAmin2_d50 network optimal triplets at filtering level 4 contain variants in five of the unsolved AOS cases. One optimal triplet comprises *LRP2*, *MAP3K11* and *MAPK8IP1*, which formed part of the network found in PINA_d50. This subnetwork has a highly significant KGGSeq-prioritisation score (probability of containing a disease-causing variant = 0.9712, $p = 2.0 \times 10^{-5}$). There are three other optimal triplets which overlap, forming a merged region comprising *MAPK1*, *NRIP1*, *RPS6KA2*, *RXRA* and *THRA* (see Figure 7.13). The quadruplet search in PINAmin2_d50 found three optimal subnetworks covering six AOS cases each, and these all included *MAPK1* and *RXRA*. Together, the triplets and quadruplets including *MAPK1* and *RXRA* form a merged region of nine genes. None of the triplets or quadruplets has a significant KGGSeq-prioritisation score, but the network region is reasonably densely connected, having three edges more than strictly necessary for connectivity. Interestingly, when a GO enrichment analysis is performed all nine genes are

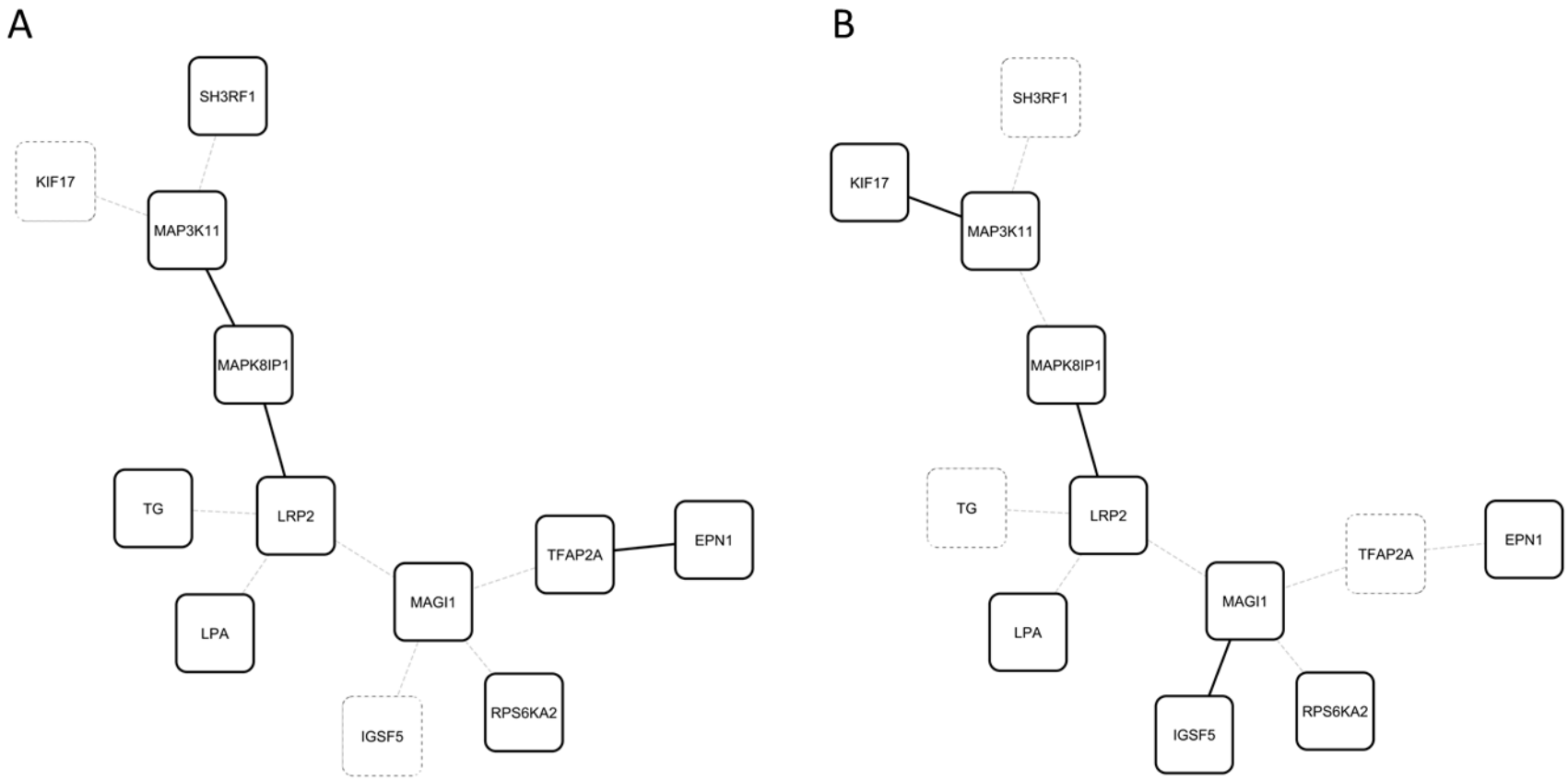
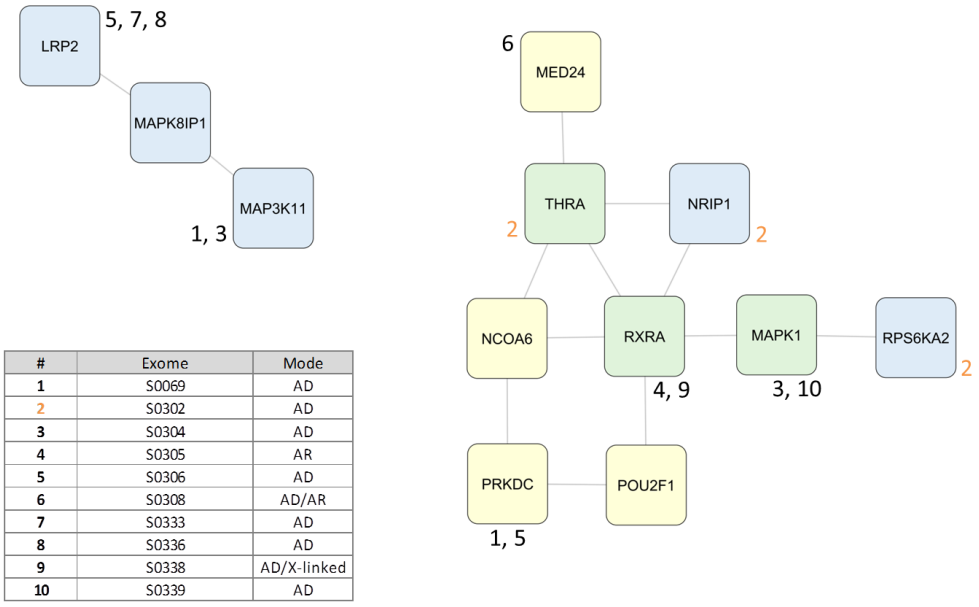


Figure 7.12 – PINA_d50 optimal AOS filtering level 4 regions retained in higher-confidence PINs
The 12 genes are those found by optimal searches in PINA_d50 network (as depicted in Figure 7.7). Bold line: node or edge retained in higher-confidence PIN; dotted line: node or edge not retained in higher-confidence PIN. (A) PINAmin2_d50; (B) CPDBconf95_d50.



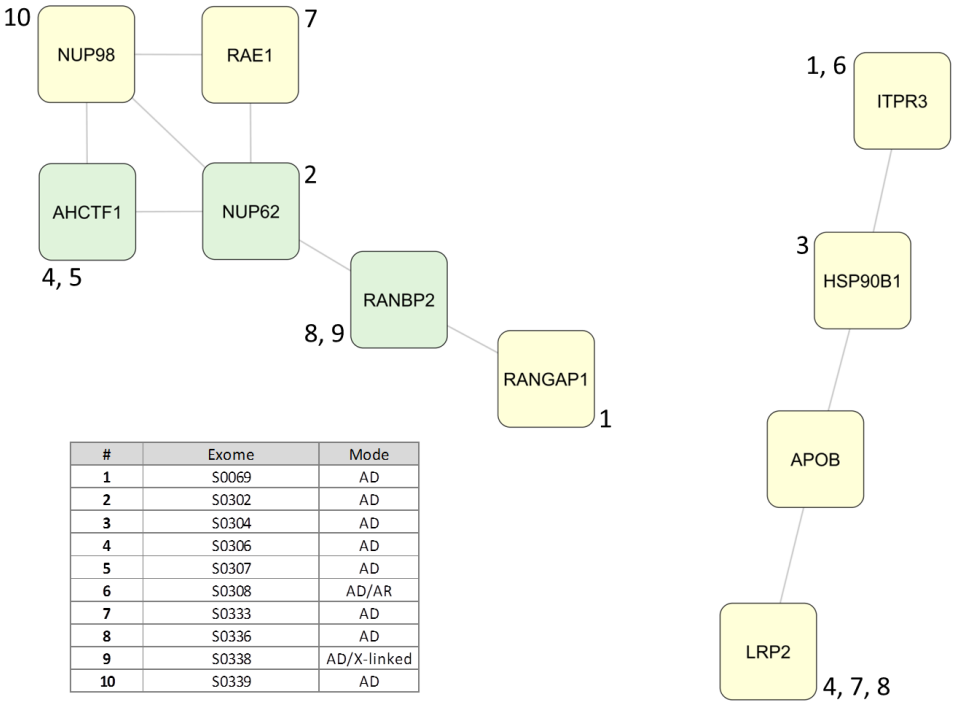
#	Exome	Mode
1	S0069	AD
2	S0302	AD
3	S0304	AD
4	S0305	AR
5	S0306	AD
6	S0308	AD/AR
7	S0333	AD
8	S0336	AD
9	S0338	AD/X-linked
10	S0339	AD

Figure 7.13 – Optimal subnetworks found in PINAmin2_d50 at AOS filtering level 4 using triplet and quadruplet searches
Merged regions shown. Genes found in optimal triplets are in blue; quadruplets in yellow; overlap in green. The number(s) next to each node refer to the exomes in which they contain variants; coloured numbers indicate exomes with multiple variants in the same merged region.

annotated with "positive regulation of gene expression" (*adjP* = 6.03×10^{-6}), although this is a relatively high-level term.

The optimal heuristic search results (both size-limited and unlimited searches) in PINAmin2_d50 also form merged regions based around *MAPK1* and *RXRA*. However, in both cases the regions contain numerous surplus variants (that is, multiple post-filtering variants for the same exome), include several connecting genes which do not contain variants, and have non-significant KGGSeq-prioritisation scores. The nine-gene region of optimal triplets and quadruplets around *MAPK1* and *RXRA*, shown in Figure 7.13, would therefore be a better candidate for follow-up study. Note there is no overlap between any of the genes identified by the heuristic searches and the region of 12 genes identified in PINA_d50 by the optimal searches at filtering level 4 (as depicted in Figure 7.7).

In the CPDBconf95_d50 network, optimal triplets again contain variants in five of the unsolved AOS cases at filtering level 4. Eight optimal triplets were found which include the gene *NINL*. Since this gene itself contains variants in four AOS cases these results are not of further interest (note that we did not find optimal triplets containing *NINL* in PINAmin2_d50 because the gene is absent from that network). The only other optimal triplet identified comprises the genes *AHCTF1*, *NUP62* and *RANBP2* (depicted in Figure 7.14), the first two of which were identified by the optimal heuristic searches and near-optimal triplet search at filtering level 1 in PINA_d50 (see section 7.3.4.4, Figure 7.10 and



#	Exome	Mode
1	S0069	AD
2	S0302	AD
3	S0304	AD
4	S0306	AD
5	S0307	AD
6	S0308	AD/AR
7	S0333	AD
8	S0336	AD
9	S0338	AD/X-linked
10	S0339	AD

Figure 7.14 – Two regions of optimal triplets and quadruplets found in CPDBconf95_d50 at AOS filtering level 4
Optimal quadruplets shown in yellow; overlap with optimal triplets in green. The number(s) next to each node refer to the exomes in which they contain variants. Not shown: network region based around the gene *NINL*.

Figure 7.11). However, none of the optimal triplets overlap with the optimal subnetworks found at filtering level 4 in PINA_d50 (Figure 7.7).

Optimal quadruplets in CPDBconf95_d50 harbour variants in six unsolved AOS cases. 12 such subnetworks contain *NINL* and are not discussed further. There are three optimal quadruplets that include *AHCTF1* and *NUP62*, and in a distinct network region another which includes *LRP2* (which was central to many of the optimal subnetworks found in PINA_d50). These findings are presented in Figure 7.14. The quadruplet which includes *LRP2* has a significant KGGSeq-prioritisation score (probability of containing a disease-causing variant = 0.8345, $p = 0.0484$).

The heuristic searches limited to subnetworks of ten genes or fewer can cover 11 of the 13 unsolved AOS exomes in CPDBconf95_d50. However, BioGranat-IG finds 35 ways to do this, resulting in a merged region of 53 genes. This region is too big to provide any specific insights into mechanisms of AOS (and it contains many surplus mutations and connecting genes containing no variants). However, it contains a number of genes we have previously seen: *NINL*; all six of the genes in the region of optimal quadruplets around *AHCTF1* and *NUP62* in Figure 7.14, and three genes (*KIF17*, *LRP2* and *MAP3K11*) that

feature in the optimal region in PINA_d50 (as depicted in Figure 7.7). We conclude that in this case the heuristic search acts so as to connect together previously-identified regions in which a concentration of post-filtering variants are found. The unlimited heuristic searches give materially similar findings so will not be described in further detail.

The heuristic searches in the higher-confidence subnetworks provide little support for the results seen in PINA_d50. This highlights a potential weakness of the minimum distance and multi-minimum distance search methods in BioGranat-IG: subnetworks will continue to be extended until all exomes are covered (i.e. carry a variant in the subnetwork); since the input data are noisy (each exome has many variants in the network) and the average path length in a network is small, it will usually be possible to find network-dependent paths between a few key genes that contain enough variants to be found by BioGranat-IG as a viable optimal subnetwork. This leads to large and difficult-to-interpret regions harbouring several post-filtering variants for many of the exomes and including several connecting genes without variants. Therefore, it is perhaps more fruitful to focus on the optimal triplets and quadruplets found by BioGranat-IG, since these tend to identify small regions more densely enriched for post-filtering variants.

7.3.5.2 Filtering levels 1-3

A full discussion of all results found at filtering levels 1-3 in the higher-confidence PINs would be too lengthy for inclusion here, but we can make several observations concerning the known AOS genes and the extent to which the PINA_d50 results are supported.

Firstly, at filtering level 3 (which is the same as level 4, except that the solved AOS cases are included with only the true causal variants) the optimal triplets and quadruplets identified in both PINAmin2_d50 and CPDBconf95_d50 are the same as those found at filtering level 4. This suggests that there are not sufficient post-filtering variants in the vicinity of any of the known AOS genes to form highly-mutated triplets or quadruplets. This is despite the fact that both higher-confidence PINs contain *ARHGAP31*, *NOTCH1* and *RBPJ* (unlike PINA_d50 in which the only known AOS gene present is *RBPJ*).

However, in PINAmin2_d50 one of the seven distinct regions formed by near-optimal triplets contains *NOTCH1* (22 genes in total, including *MAPK1* and *RXRA*). One of the three distinct regions formed by near-optimal quadruplets contains both *NOTCH1* and *RBPJ* (63 genes in total, including *MAPK1* and *RXRA* as well as several genes from the optimal heuristic subnetwork in PINA_d50 at filtering level 3: *LRP2*, *MAP3K11*, *MAPK8IP1* and *RPS6KA2*). In CPDBconf95_d50, one of the three distinct regions formed by near-optimal quadruplets also contains both *NOTCH1* and *RBPJ* (58 genes in total,

including *AHCTF1* and *NUP62*, as well as *MAP3K11* from the optimal heuristic subnetwork in PINA_d50).

The heuristic searches limited to ten genes in PINAmin2_d50 find *NOTCH1* and *RBPJ*. There is no overlap (other than *RBPJ*) with any optimal subnetwork genes found in PINA_d50 at filtering level 3. The unlimited heuristic searches find *ARHGAP31*, *NOTCH1* and *RBPJ* in a network region of 31 genes in total, which includes *EPN1* and *TFAP2A* from the PINA_d50 results. Similarly, the size-limited heuristic searches in CPDBconf95_d50 find *NOTCH1* and *RBPJ* and there is no other overlap with optimal subnetwork genes found in PINA_d50, while the unlimited searches find *ARHGAP31*, *NOTCH1* and *RBPJ* in a merged region of 27 genes, with *MAP3K11* the only gene overlapping with PINA_d50 results. As with filtering level 4, the heuristic searches in the higher-confidence PINs at filtering level 3 provide little support for the results in PINA_d50, again suggesting that the most relevant BioGranat-IG results will come from the triplet and quadruplet searches.

Secondly, at filtering level 2 (which is the same as level 3, except that exomes for the solved cases now include all variants matching the expected mode of inheritance, and not just the true causal variants) there are five distinct network regions formed by optimal triplets in PINAmin2_d50. One includes four genes (*LRP2*, *MAP3K11*, *MAPK8IP1* and *MYO6*) that were found in the PINA_d50 filtering level 2 optimal subnetworks (although the latter three were only found in the heuristic search results) and another includes the gene *RPS6KA2*, which was in an optimal quadruplet in PINA_d50. Of the three regions formed by optimal quadruplets, one includes *DAB1*, *LRP2* and *RELN*, all of which were found in optimal quadruplets in PINA_d50. Notably, with the extra (non-causal) variants that are introduced at filtering level 2 for the solved cases, none of the known AOS genes are present in any of the optimal subnetworks in PINAmin2_d50.

In CPDBconf95_d50 none of the optimal triplets overlap with any of the PINA_d50 optimal subnetworks, although the unique optimal quadruplet found contains *LRP2*. Again none of the known AOS genes are present in any of the CPDBconf95_d50 optimal triplets.

Finally, at filtering level 1 (which is the same as level 2, except that variants are not filtered based on mode of inheritance) none of the optimal subnetworks found in PINAmin2_d50 (by any search method) overlap with those found in PINA_d50, and none of the known AOS genes are identified. However, in CPDBconf95_d50 there are three distinct regions of optimal triplets, one of which includes *NINL* (there were regions of optimal triplets and optimal quadruplets around this gene in PINA_d50) and one of which includes *AHCTF1* and *NUP62* (which were found by the heuristic searches in PINA_d50). The one merged region formed by optimal quadruplets also contains the latter two genes. Again,

however, none of the known AOS genes were identified in any of the optimal subnetworks in CPDBconf95_d50.

7.3.6 BioGranat-IG Results: Top Prioritised Results in all Networks

Having considered BioGranat-IG results from the PINs in detail, we can look at the KGGSeq-prioritisation p-values across all networks to establish whether there were subnetworks of interest found in the other networks.

Table 7.8 lists the prioritisation scores that achieve nominal significance ($p \leq 0.05$) for each distinct subnetwork (triplet and quadruplet searches) or merged region (all search methods) for all networks and variant-filtering levels. Of the 19 distinct subnetworks listed, 18 include the gene *LRP2*, while *LPA* (13 subnetworks), *MAGI1* (13), *MAP3K11* (six) and *MAPK8IP1* (five) are also well-represented.

LRP2 features prominently due to variants with KGGSeq probabilities of being disease-causing of 0.69495 in exome S0333, plus 0.21881 in exome S0306 and 0.09828 in exome S0336. The first two of these in particular represent high KGGSeq probabilities, as can be seen in the cumulative frequency plot given in Figure 7.15. *LPA* has variants with disease-causing probabilities between 0.05 and 0.1 in four exomes (S0308, S0338, S0038 and S0333). *MAGI1* has one between 0.05 and 0.1, and two between 0.025 and 0.05. *MAP3K11* has a variant with disease probability 0.86036 in exome S0304 (the fifth-highest probability among all post-filtering variants at level 4) and 0.03938 in exome S0069. *MAPK8IP1* does not harbour post-filtering variants itself, but connects *LRP2* and *MAP3K11* in PINA_d50 and PINAmin2_d50.

We also see from the table that 17 of the 19 subnetworks were identified in PINA_d50, PINAmin2_d50 or CPDBconf95_d50, and have been previously discussed. The optimal subnetwork with the most significant KGGSeq-prioritisation score is the *LRP2-MAP3K11-MAPK8IP1* triplet found in PINAmin2_d50 at filtering levels 4 (discussed in 7.3.5.1), 3 and 2.

In row 11 of Table 7.8 is an optimal triplet identified in Multinet_d50 at filtering levels 3 and 4 (and since no set of four genes was found containing variants in more AOS cases, the three genes also represent the best subnetwork found using the quadruplet search). However, this is the same *LPA-LRP2-MAGI1* triplet found in PINA_d50 at filtering levels 3 and 4 (row 4 in the table), and it has the same connecting edges in Multinet_d50. The same KGGSeq-prioritisation score is assigned because the same variants are observed in the same exomes (probability of containing a disease-causing variant = 0.8361), but a slightly different p-value is estimated in Multinet_d50 ($p = 0.0380$, compared to 0.0221 in PINA_d50) because permutation occurs over all variants that map into the network.

Table 7.8 – Top BioGranat-IG optimal subnetworks by KGGSeq-prioritisation score for AOS
Includes any optimal subnetwork (for any network, variant-filtering level and search method) with nominally significant KGGSeq-prioritisation score. Ordered by KGGSeq-prioritisation p-value. Table continues onto next page.

	Filtering level	Network	Search method	# Genes	# AOS cases	Prob. disease causing	p-value	Genes
1	2/3/4	PINAmin2_d50	Triplet	3	5	0.971175	0.00002	*LRP2*, *MAP3K11*, *MAPK8IP1*
2	3	PINA_d50	Heuristic – unlimited	12	14	0.984823	0.00473	*EPN1*, *FHL1*, *HIVEP3*, *LPA*, *LRP2*, *MAGI1*, *MAP3K11*, *MAPK8IP1*, *RBPJ*, *SH3RF1*, *TFAP2A*, *WWP1*
3	3/4	PINA_d50	Heuristic – limit 10 (3, 4), unlimited (4)	10	13	0.985712	0.00497	*EPN1*, *KIF17*, *LPA*, *LRP2*, *MAGI1*, *MAP3K11*, *MAPK8IP1*, *RPS6KA2*, *SH3RF1*, *TFAP2A*
4	3/4	PINA_d50	Triplet	3	7	0.836116	0.02208	*LPA*, *LRP2*, *MAGI1*
5	2	PINA_d50	Heuristic – limit 10	17	18	0.994622	0.02274	*COL6A3*, *DAB1*, *DGKD*, *DYSF*, *FLNC*, *KIF17*, *LPA*, *LRP2*, *MAGI1*, *MAP3K11*, *MAP3K12*, *MAPK8IP1*, *MBIP*, *MYO6*, *NINL*, *RPS6KA2*, *SH3RF1*
6	3/4	PINA_d50	Quadruplet	4	8	0.842374	0.02799	*IGSF5*, *LPA*, *LRP2*, *MAGI1*
7	2	PINAmin2_d50	Triplet	5	7	0.982985	0.02864	*GIPC1*, *LRP2*, *MAP3K11*, *MAPK8IP1*, *MYO6*
8	2	PINAmin2_d50	Triplet	3	5	0.873156	0.02909	*GIPC1*, *LRP2*, *MYO6*
9	3/4	PINA_d50	Quadruplet	4	8	0.843510	0.02935	*LPA*, *LRP2*, *MAGI1*, *TFAP2A*
10	2	PINAmin2_d50	Quadruplet	4/5	7	0.827065	0.03428	*DAB1*, *LRP2*, *RELN*, with *LRP8* or *VLDLR*
11	3/4	Multinet_d50	Triplet/ Quadruplet	3	7	0.836116	0.03795	*LPA*, *LRP2*, *MAGI1*
12	2	PINA_d50	Triplet	3	9	0.852532	0.03928	*LPA*, *LRP2*, *MAGI1*

Table 7.8 – Top BioGranat-IG optimal subnetworks by KGGSeq-prioritisation score for AOS (continued)

	Filtering level	Network	Search method	# Genes	# AOS cases	Prob. disease causing	p-value	Genes
13	1	COXPRES30_d50	Quadruplet	4	10	0.936203	0.04250	*ABCA7*, *ARHGAP4*, *MAP3K11*, *SIPA1*
14	2	PINA_d50	Quadruplet	4	10	0.860602	0.04487	*DAB1*, *LPA*, *LRP2*, *MAGI1*
15	3/4	PINA_d50	Quadruplet	4	8	0.841958	0.04495	*LPA*, *LRP2*, *MAGI1*, *TG*
16	2	PINA_d50	Quadruplet	4	10	0.859761	0.04550	*LPA*, *LRP2*, *MAGI1*, *PIP5K1C*
17	2	PINA_d50	Quadruplet	4	10	0.858163	0.04628	*IGSF5*, *LPA*, *LRP2*, *MAGI1*
18	3/4	PINA_d50	Quadruplet	4	8	0.837872	0.04817	*LPA*, *LRP2*, *MAGI1*, *RPS6KA2*
19	3/4	CPDBconf95_d50	Quadruplet	4	6	0.834526	0.04839	*APOB*, *HSP90B1*, *ITPR3*, *LRP2*

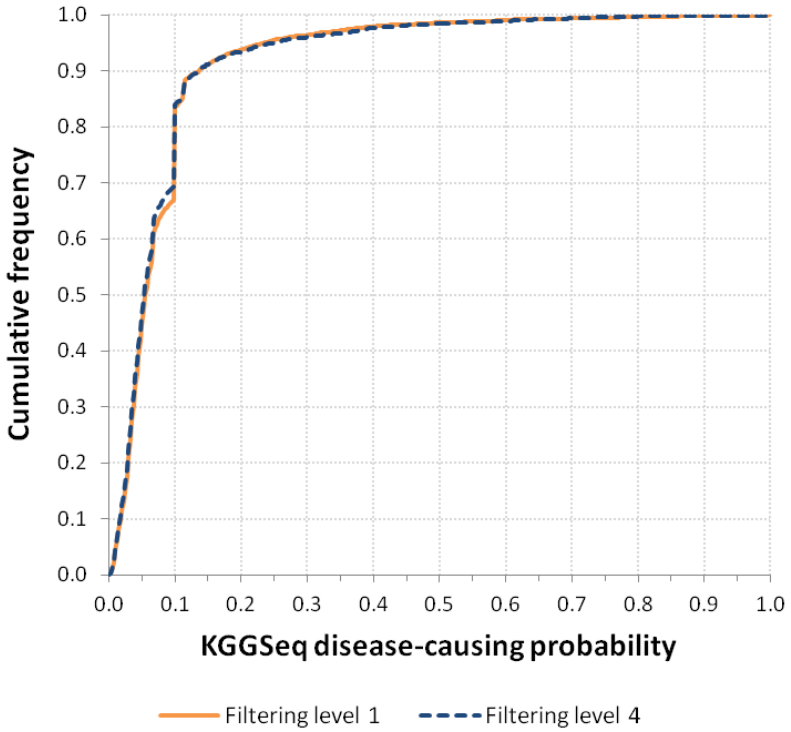


Figure 7.15 – Cumulative frequency plots of probabilities for causing disease that KGGSeq assigns to post-filtering variants in AOS exomes
Cumulative frequencies taken across all post-filtering variants at filtering level 1 (across all 19 AOS exomes) and filtering level 4 (13 unsolved exomes only) are highly similar.

The only subnetwork in COXPRES30_d50 that has a significant KGGSeq-prioritisation score is an optimal quadruplet identified at filtering level 1 (row 13 in Table 7.8; probability of containing a disease-causing variant = 0.9362, $p = 0.0425$). It comprises the genes *ABCA7*, *ARHGAP4*, *MAP3K11* and *SIPA1* and contains variants for 10 of the 19 AOS exomes (see Figure 7.16). The gene *ARHGAP4* stands out because it encodes a Rho GTPase-activating protein (as does the AOS-causing gene *ARHGAP31*), and contains a variant in four of the 19 AOS exomes (which is also true of its neighbour, *ABCA7*; both genes are listed in Table 7.5 due to the high burden of post-filtering variants they contain at level 1). However, two of the variants that *ARHGAP4* contains are in solved AOS cases: S0311 is known to be caused by *NOTCH1* and S0337 by *RBPJ*. Further, the variant that *ARHGAP4* harbours in exome S0302 is homozygous, but the expected mode of inheritance for this case is AD. Considering these observations, the fact that two of the exomes in which *ABCA7* contains a variant are also solved cases for which other genes cause AOS, and the fact that the subnetwork is not densely connected, contains one surplus variant and one connecting gene with no post-filtering variant, there is no strong argument for further investigation of these genes.

#	Exome
1	S0069
2	S0302
3	S0304
4	S0307
5	S0308
6	***S0311***
7	***S0334***
8	***S0335***
9	***S0337***
10	S0338

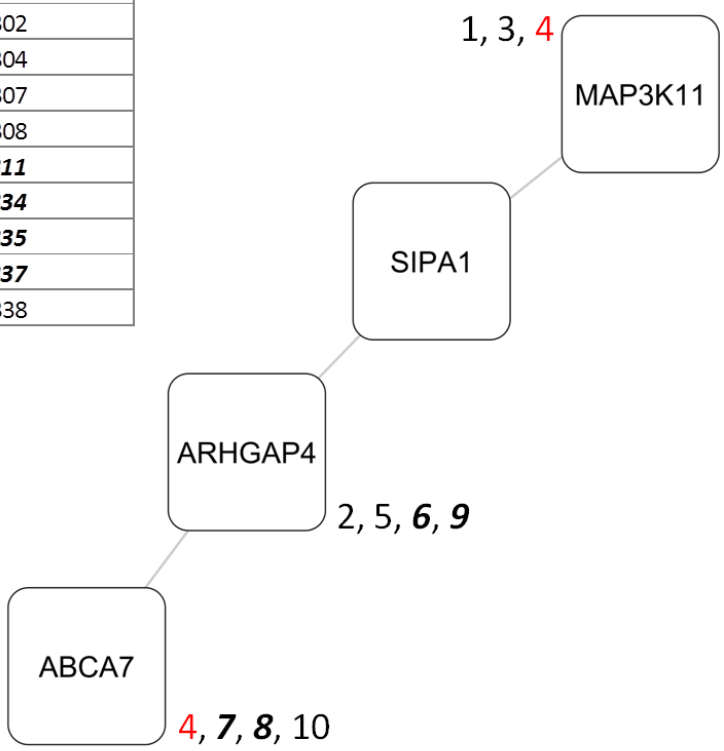


Figure 7.16 – KGGSeq-significant optimal quadruplet found in COXPRES30_d50 at AOS filtering level 1
The numbers next to each node refer to the exomes in which they contain variants; bold italic indicates a solved AOS case with causal mutations found elsewhere; coloured numbers indicate the exome has multiple variants in the region.

7.3.7 HetRank Results: PINA Network

While BioGranat-IG outputs specific variant-harbouring subnetworks, HetRank ranks all genes according to the evidence that they contain an AOS-causing variant (with network neighbours playing an important role in the ranking). Therefore a different approach is needed to interpret HetRank results and identify plausible functional pathways that are enriched for highly-ranked genes (and thus could underlie AOS). In this section, results from the full PINA network will be discussed; subsequent sections will compare these results to those found in the other networks and examine how the results change as a result of minor changes to the input data and parameters (that are intended to better discriminate causal variants).

HetRank does not require pre-filtering of variants, and is therefore not performed using different levels of filtering criteria. However, in considering the PINA results, we will first look at the results of employing HetRank analysis using only the 13 unsolved exomes (as with BioGranat-IG this allows the possibility of identifying novel AOS pathways without

Table 7.9 – Top 20 genes ranked by HetRank without network-based rank adjustment, based on 13 unsolved AOS exomes

Rank	Gene	In PINA network?
1	*ABCA10*	✓
2	*ZDHHC13*	✓
3	*EXO5*	×
4	*NINL*	✓
5	*PRAMEF1*	×
6	*SLC9B1*	×
7	*COTL1*	✓
8	*VCX*	×
9	*GJB2*	✓
10	*RGPD3*	✓

Rank	Gene	In PINA network?
11	*CCDC144NL*	×
12	*AGAP6*	×
13	*NBPF11,NBPF24*	×
14	*GNRH2*	✓
15	*NBPF15,NBPF16*	×
16	*LAPTM4B*	✓
17	*ADAMTS7*	×
18	*FAM86B2*	✓
19	*CNTNAP3B*	✓
20	*NBPF6*	×

a close functional relationship to known AOS genes), and then compare this to the results based on all 19 AOS exomes (which also allows us to judge how well HetRank picks out the known AOS genes).

7.3.7.1 HetRank Analysis Using the 13 Unsolved AOS Exomes

The HetRank tool works by first ranking genes in each exome according to the evidence that they contain a disease-causing variant, before adjusting the rankings based on neighbouring genes in the network and producing a final rank by combining across exomes. Therefore a logical first step is to consider the results obtained without performing the network-based adjustment, which should indicate the genes that carry most evidence independently of their network neighbours.

Table 7.9 gives the genes ranked 1-20 without network-adjusted rankings, also indicating whether each gene is present in the PINA network. These final positions combine separate rankings across all 13 unsolved exomes; to understand where the evidence supporting these genes comes from we can examine the individual exomes in which each gene receives a high ranking. When variants were filtered (see Table 7.2 in the Methods section of this chapter) we saw that the number of plausible disease-causing variants in a typical exome was of the order of 200. This gives a convenient threshold with which to address the rankings in each individual exome.

ABCA10 is ranked first overall. Four of the 13 unsolved AOS exomes have *ABCA10* ranked among the top 200 genes likely to contain the disease-causing variant. In S0040 there are two heterozygous frameshift deletions that are not present in 1000 Genomes or EVS data (although each has been seen 24 times in heterozygous form in the in-house exome

database); S0302 and S0336 also contain these same two frameshift deletions, and S0338 contains a different heterozygous frameshift deletion that is not present in 1000 Genomes or EVS data but has been seen 15 times in the in-house exome database. These variants result in a high rank for *ABCA10* in four of the exomes because a low weight is assigned to the ranking factor "number of observations in heterozygous form in the in-house exome database". However, in practice the previous observations (in non-AOS exomes) rule out the possibility that these variants cause AOS. While one of the aims of the HetRank tool is to dispense with the need for fixed filtering thresholds for exome data, this is an example of a case where a simple filter would have removed these variants from further consideration. Without resorting to fixed filtering thresholds, one option available is to increase the weight assigned to this ranking factor. This will be explored further in section 7.3.9.2.

ZDHHC13 is ranked second overall, also having four exomes in which it ranks among the top 200 genes. In S0305 it contains a novel heterozygous frameshift insertion; in S0308 it contains a heterozygous missense SNV that is not present in 1000 Genomes or EVS data but has been seen once in heterozygous form in the in-house exome database; in S0333 there is a novel heterozygous missense SNV, and in S0339 there is a heterozygous missense SNV which is not present in 1000 Genomes or EVS data and is not in the in-house exome database (but is not strictly novel as it has a dbSNP reference number – note that presence or absence of a dbSNP identifier is not used as a ranking factor in HetRank). Inspection of these variants would suggest that *ZDHHC13* is a better candidate than *ABCA10* to be involved in an AOS disease mechanism. (Note that the four variants described were all present after filtering at level 1, the only one filtered out at levels 2-4 being the insertion in S0305, since this individual is expected to have an AR form of AOS caused by a homozygous mutation. This is reflected in Table 7.5.) *ZDHHC13* encodes a palmitoyltransferase (which is not inconsistent with a signalling role) and has been associated with a wide range of phenotypes in mouse models (www.genecards.org, Stelzer et al. 2011).

EXO5 is ranked third overall, with four exomes in which it ranks among the top 200 genes. It contains a low-frequency heterozygous missense SNV in the shared S0039/S0301 exome and identical low-frequency heterozygous frameshift insertions in S0306, S0308 and S0339 (albeit a variant that has been seen 21 times in the in-house exome database in heterozygous form). As with *ABCA10* the previous observations of the frameshift insertion mean we have little evidence to support an AOS role for *EXO5*.

It is of course possible to consider every gene in this list individually. However, we will just note that the other genes ranked in the top 20 all have three or fewer top-200 rankings in individual exomes, and that there are some genes here that we have seen

previously. In particular, *NINL* (ranked 4th) and *CNTNAP3B* (19th) were picked out in Table 7.5 because they contained the highest number of post-filtering variants at various filtering levels. *RGPD3* (10th), *AGAP6* (12th) and, as previously mentioned, *ZDHHC13* (2nd) also featured in Table 7.5 due to their burden of post-filtering variants; this is reassuring because ranking and filtering provide different approaches to the same problem, and hence some overlap is expected.

By considering the results of performing HetRank analysis using the PINA network, we can see how these rankings change when evidence that a gene is involved in AOS is supported by its network neighbours; if AOS is caused by several functionally-related genes in these 13 cases we would hope that PINA reflects this relationship and hence the ranking of these genes is improved. Table 7.10 gives the results of performing HetRank analysis on the 13 unsolved AOS cases using the PINA network.

The top-ranked gene is *POLN*. This gene already had a relatively high ranking of 25 (out of 18,838 genes in total) in the non-network-adjusted rankings due in part to top-200 rankings in three of the exomes. In S0302 *POLN* harbours a low-frequency heterozygous frameshift deletion (seen twice in the in-house exome database) but its pre-network ranking of 20 in this exome falls to 71 after the network adjustment as a result of other genes having their rankings improved elsewhere in the network. Interestingly, the other two exomes in which *POLN* has a pre-network ranking in the top 200 still have a ranking in the top 200 after network adjustment, but due in part to better-ranked variants in network neighbours.

In the shared S0039/S0301 exome *POLN* has a pre-network rank of 170 thanks to a low-frequency missense SNV. However, *POLN* is a direct neighbour in PINA of the gene *ATR*, which was ranked joint-top in the pre-network ranking for this exome due to a novel frameshift deletion. HetRank works on the premise that this connection in the network could signify a shared functional relationship; *POLN*'s rank is therefore adjusted to reflect this and it is consequently ranked third in this exome. (The ranking adjustment takes neighbourhood size into account, but since the direct neighbourhood of *POLN* contains only nine genes, compared to a network maximum of 7,805 genes, the adjusted ranking is almost fully weighted toward *ATR*'s value. Reassuringly, since there are no genes in the network with a better rank, *ATR*'s rank is not adjusted and it remains joint-top in the final rankings for this exome.)

Likewise in exome S0307 *POLN* has a pre-network rank of 104 due to a novel heterozygous missense SNV. However, *POLN* is an indirect neighbour in PINA of the gene *ANTXR2*, which has a pre-network rank of 28 in this exome (also due to a novel heterozygous missense SNV; the difference in ranking must reflect the fact that the ranking factor derived from non-AOS control exomes rates this occurrence as more unusual in

Table 7.10 – Top 20 genes ranked by HetRank using PINA network, based on 13 unsolved AOS exomes
"Rank" = overall ranking combined across all 13 exomes after network-based adjustment; "# top 200 ranks" = number of the 13 exomes in which gene is ranked ≤200; "Pre-network rank" = overall ranking combined across all 13 exomes before network-based adjustment (the top 20 such genes were shown in Table 7.9); N_d = gene's *d*-neighbourhood size (number of genes within distance *d*).

Rank	Gene	# top 200 ranks	Pre-network rank	# top 200 ranks (pre-)	N_1	N_2
1	*POLN*	3	25	3	9	353
2	*FCGR3A*	3	224	0	14	412
3	*KIF17*	3	319	1	8	256
4	*SH2D1B*	3	846	1	7	165
5	*ABCA10*	4	1	4	3	770
6	*HGFAC*	2	1,510	0	7	111
7	*LPA*	4	98	1	13	524
8	*DBH*	2	779	0	3	98
9	*GLYCTK*	2	7,420	0	4	69
10	*FRZB*	2	728	0	6	153
11	*IL12RB2*	3	385	1	7	161
12	*FXYD3*	3	1,067	2	6	306
13	*CINP*	2	1,663	1	4	90
14	*BCR*	2	75	2	71	9,129
15	*ADCY2*	2	6,026	0	5	104
16	*SLC26A8*	1	4,983	0	4	144
17	*MST1*	1	22	1	6	88
18	*SFTPC*	1	1,460	1	5	293
19	*MUC4*	3	2,963	0	3	131
20	*SOX7*	2	72	2	3	95

ANTXR2). The final ranking of *POLN* in S0307 is thus adjusted based on *ANTXR2*'s score (although adjustments to other genes mean *POLN*'s final ranking in this exome actually falls to 117).

This illustrates an unexpected consequence of HetRank's network-based rank adjustment. In only one of the 13 unsolved AOS exomes (S0302) is *POLN*'s rank not adjusted because of a better-ranked neighbour. This is not exceptional: of the top 20 genes in Table 7.10, 14 have their pre-network rankings adjusted due to a better-ranked neighbour in all 13 exomes. The gene whose ranking is least-frequently adjusted is *ABCA10* (which was ranked top overall in the non-network-adjusted rankings), and even this remains unadjusted in only three of the 13 exomes.

This feature of the rank adjustment (that very few ranks remain unadjusted) is explained by the typical neighbourhood sizes of genes in the network: only one member of a gene's neighbourhood is required to have a better rank for an adjustment to be made, and

most neighbourhoods contain many genes (most of the 2-neighbourhoods in Table 7.10, for example, contain hundreds of genes). If AOS has a basis of locus heterogeneity which can be understood in terms of a functional relationship in PINA (and if the 13 unsolved AOS cases reflect this underlying relationship), we would probably expect to see the genes ranked highest by HetRank having high ranks (around the top 200) in several exomes due to variants they themselves contain, and high ranks in several other exomes due to variants in a handful of neighbours. The fact that we are not seeing this (our top 20 genes are being adjusted in all or nearly all exomes, and no gene has a top-200 ranking in more than four exomes after adjustment) could indicate a flaw in the design of HetRank, or alternatively that there is no underlying functional pathway in PINA that causes AOS for these cases.

The gene ranked second overall is *FCGR3A*, which had a non-network-adjusted rank of 224. In all 13 exomes *FCGR3A*'s rank is adjusted due to a better-ranked neighbour; in three exomes this results in a top-200 ranking: in S0302 its rank is adjusted due to a novel heterozygous nonsense SNV in the indirect neighbour *PIK3C2B* (ranked joint-top pre-adjustment in that exome); in S0308 its rank is adjusted due to a novel heterozygous nonsense SNV in the indirect neighbour *LAT* (ranked joint-top pre-adjustment), and in S0333 its rank is adjusted due to a novel heterozygous nonsense SNV in the indirect neighbour *GRAP2* (ranked joint-top pre-adjustment). *FCGR3A* has 398 indirect neighbours, and the fact that three of these contain severe variants in AOS exomes is insufficient grounds to conclude that this gene plays a role in AOS.

KIF17 is ranked third overall, compared to a non-network-adjusted ranking of 319. This is an interesting gene because we have previously seen it in the optimal subnetworks identified by BioGranat-IG at filtering level 4 in the PINA_d50 network (see Figure 7.7). There, *KIF17* had a post-filtering variant in S0302 while its direct neighbour *MAP3K11* contained variants in S0069 and S0304, and its indirect neighbour *SH3RF1* contained a variant in the shared S0039/S0301 exome.

What happens to *KIF17*'s ranking in these exomes in HetRank? The novel heterozygous missense SNV that *KIF17* contains in S0302 gives it a rank of 91 before adjustment; however *KIF17* is an indirect neighbour of *NUP62* (another gene we recognise from the PINA_d50 BioGranat-IG results – see Figure 7.10 and Figure 7.11), which is ranked 36.5 before adjustment due to its own novel heterozygous missense SNV. *KIF17*'s rank is therefore adjusted to take into account *NUP62*'s score, giving a final rank of 184 in exome S0302 (which is lower than either gene's pre-adjustment rank due to other genes having their rankings improved elsewhere in the network). In S0069, the novel heterozygous nonsense SNV in *MAP3K11* does indeed boost *KIF17*'s ranking from 2,756 to 10 and in S0304 the novel heterozygous missense SNV in *MAP3K11* boosts *KIF17*'s ranking from

411 to 222. Finally, in the shared S0039/S0301 exome *KIF17*'s rank is adjusted, but not by *SH3RF1* which we saw in the BioGranat-IG results. While *SH3RF1* harbours a low-frequency heterozygous missense SNV which gives it a rank of 104 in this exome before adjustment, another indirect neighbour, *XRCC1*, has a pre-adjustment rank of 91 due to a rarer missense SNV. However, this is only enough to boost *KIF17*'s rank from 4,537 to 2,826 in this exome. There is one other exome in which *KIF17* has a top-200 rank after adjustment. In exome S0339 *KIF17* has a pre-adjustment rank of 271, but its indirect neighbour *KALRN* has a rank of 10 due to a low-frequency heterozygous nonsense SNV. This results in an adjusted rank of 56 for *KIF17*.

Some of these examples of how *KIF17*'s ranks are adjusted in individual exomes hint at another subtlety of HetRank's network-based rank-adjustment process. When a network neighbour has a better rank in a given exome, HetRank adjusts the score by which a gene will be ranked. The new score is a weighted average of the gene's original score and its neighbour's original score; the weight is based on neighbourhood size and reflects the extent to which a better-ranking neighbour is assumed to be relevant (genes with very large neighbourhoods receive only small adjustments to their scores because it is more likely they will have a high-ranking neighbour by chance). While this formulation reflects a level of belief about the value of the information obtained by looking at network neighbours, in practice the consequence is that a gene's adjusted rank depends on both it and its neighbour's original scores (when under the assumption that only one variant in each exome can cause AOS, one of these should have no disease role). However, the relationship is complicated by the weight derived from the neighbourhood size. Figure 7.17 shows that genes with high original rank tend to have a relatively high rank after adjustment (that is, higher than genes with lower original rank that have a neighbourhood of similar size).

We see several different types of adjustment for *KIF17*: for example, its rank is boosted from 2,756 to 10 in S0069 (relatively low original rank but the direct neighbour *MAP3K11* has a particularly high rank and the direct neighbourhood, at eight genes, is small); in the shared S0039/S0301 exome its rank is only boosted from 4,537 to 2,826 (low original rank, the indirect neighbour *XRCC1* has a relatively good rank of 91 but the indirect neighbourhood has 256 genes), and in S0339 its rank is boosted from 271 to 56 (relatively high original rank, the indirect neighbour *KALRN* has a high original rank and the neighbourhood has 256 genes).

There are a few other genes of note in Table 7.10. As previously noted, *ABCA10* (the top-ranked gene before the network-based adjustment) is ranked fifth.

Ranked seventh is *LPA*, which was a key gene in the optimal subnetworks identified by BioGranat-IG at filtering level 4 in the PINA_d50 network (see Figure 7.7). There, *LPA*

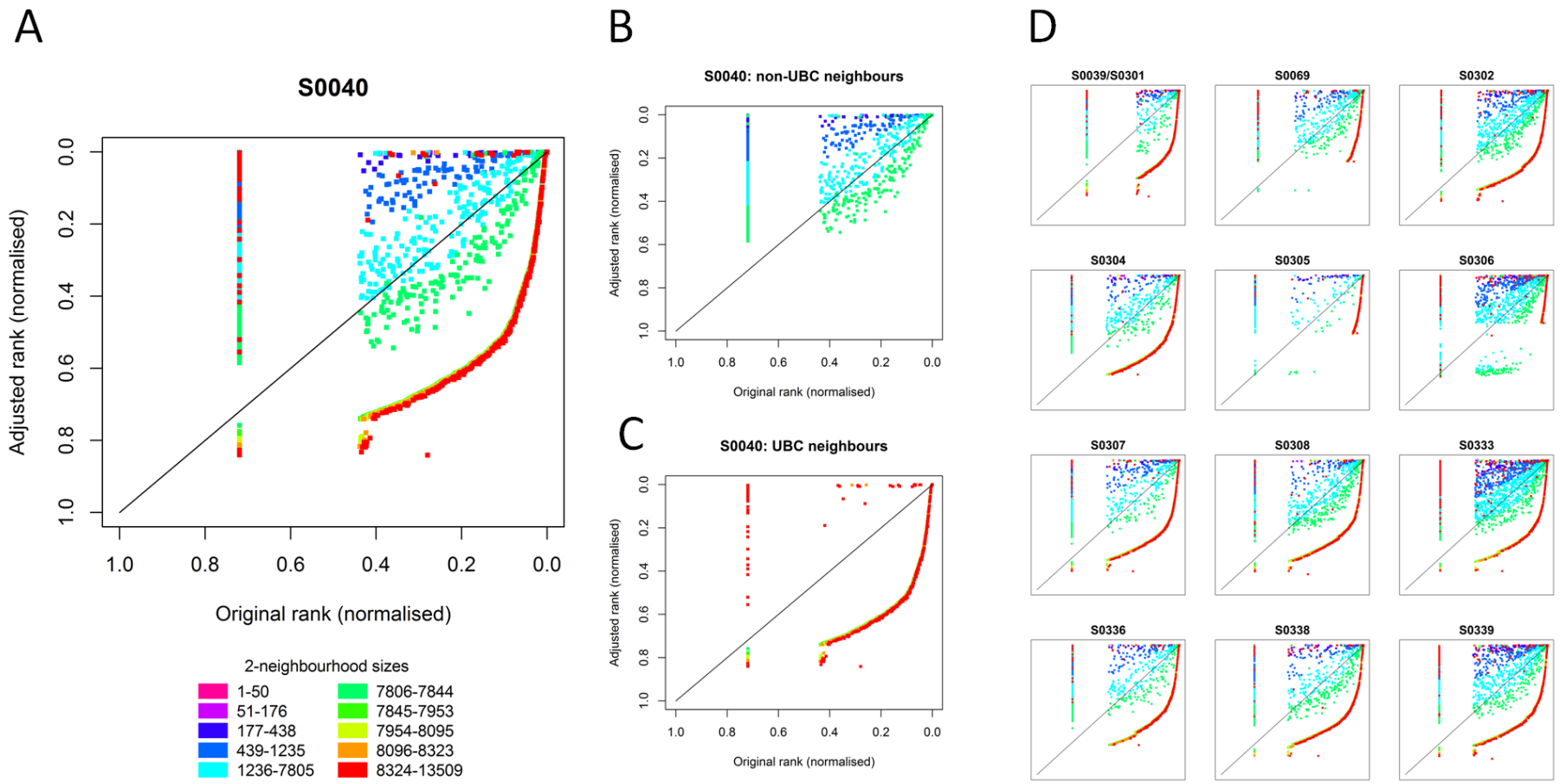


Figure 7.17 – Relationship between original rank, adjusted rank and neighbourhood size, by exome
See next page for full figure legend.

Figure 7.17 – Relationship between original rank, adjusted rank and neighbourhood size, by exome (previous page)
(A) Original vs. adjusted rankings in exome S0040, for genes whose ranks are adjusted due to a neighbour with original rank of 20 or less. Ranks are normalised to range (0,1] and best-ranked genes are at top/right of plot. Each point represents a gene; colour represents the size of the gene's 2-neighbourhood (ranges chosen to give approximately equal number of genes in the network of each colour). The vertical line at 0.719 represents genes not containing a variant, which are all ranked joint last before network adjustment. Some stratification by neighbourhood size is evident (genes with smaller neighbourhoods [e.g. purple to blue] tend to have their ranks improved more than genes with bigger neighbourhoods [e.g. green to red]; this is expected because HetRank assumes high-ranking neighbours are more relevant when the neighbourhood is small). Also evident is that adjusted rank is not independent of original rank: genes with high original rank tend to have a relatively high adjusted rank (more so than lower-ranked genes with similar network size, i.e. of the same colour). The curve of mainly red, orange and yellow points represents genes with a large neighbourhood size; although they do have a neighbour with original rank of 20 or less, their adjusted rank is low because little weight is given to this neighbour's rank. The shape of this curve demonstrates how the adjusted rank falls as a result of other genes in the network having their ranks boosted. (B) and (C) demonstrate that the red/orange/yellow curve is entirely explained by genes of large neighbourhood size: PINA has one gene of exceptionally high degree (*UBC*); genes which are direct neighbours of *UBC* necessarily have a very large 2-neighbourhood. (B) and (C) show the same information as (A) but for the subset of genes that are not/are direct neighbours of *UBC*, respectively. (D) Plots for the other 12 unsolved AOS exomes show that the relationships in S0040 are typical.

had a post-filtering variant in three exomes (S0308, S0333 and S0338). In HetRank, *LPA*'s rank is adjusted in all three of these exomes: in S0308 it is originally ranked 166 but has an indirect neighbour (*PIGK*) ranked 32.5; in S0333 it is originally ranked 283 but has an indirect neighbour (*BCR*) ranked 20, and in S0338 it is originally ranked 667.5 but has an indirect neighbour (*TAC1*) ranked 25.5. In the BioGranat-IG results *LPA*'s direct neighbour *LRP2* also had a post-filtering variant in three exomes (S0306, S0333 again and S0336) and its indirect neighbours *TG* (S0304) and *MAGI1* (S0040 and S0305) also contained post-filtering variants. In all of these exomes *LPA*'s rank is adjusted not because of these genes but because of other better-ranked neighbours. A key difference between HetRank and BioGranat-IG is that while BioGranat-IG searches for optimal subnetworks with respect to the post-filtering variants in all exomes together, HetRank is free to adjust a gene's rank in any exome by examining all of its direct and indirect neighbours – independently of the variants in other exomes.

Ranked ninth is *GLYCTK*, of note due to its pre-adjustment rank of 7,420. This change comes about because *GLYCTK* is a direct neighbour of the gene ranked second in S0338 and an indirect neighbour of the genes ranked 4.5 in the shared S0039/S0301 exome, 4.5 in S0306 and 12 in S0339 (so that it ranks in the top 500 for four exomes post-adjustment, two being top-200 rankings). Furthermore, *GLYCTK* has relatively small direct (4 genes) and indirect (69 genes) neighbourhoods, which means that the weighting used in the adjustment favours the neighbours' scores. This serves to illustrate how volatile the adjustment to gene rankings can be.

Looking at individual genes that are highly ranked by HetRank has allowed us to study the evidence supporting a role in AOS for each gene, but it is difficult to draw any clear conclusions about an underlying disease mechanism in this way. The RGA tool (Lehne 2011) offers a convenient way to systematically pick out network regions that are enriched for genes ranked highly by HetRank, thereby proposing candidate functional pathways linked to AOS.

RGA was employed to test all alpha thresholds in the range $1 \leq \alpha \leq 250$, excluding outlier genes (so that for each value of alpha RGA identifies the largest connected network region comprising genes with rank less than or equal to alpha), and with empirical p-values determined with reference to 10,000 degree-constrained permuted networks (the significance of an identified region in the original network is estimated by the proportion of permuted networks in which a region of equal or greater size is found). Region sizes and p-values are presented in Figure 7.18a. There are three ranges of alpha values at which we see a nominally significant region size: $\alpha = 21$ (region size 2), $24 \leq \alpha \leq 65$ (region size 3, increasing to 4 at $\alpha = 61$) and $148 \leq \alpha \leq 151$ (region size 7). The lowest p-value is seen at $\alpha = 24$ ($p = 0.0015$). The regions identified are presented in Figure 7.18b. Note that three distinct network regions are found at different alpha levels; note also that none of the regions support any of the main subnetworks found by BioGranat-IG in the PINA_d50 network (discussed in section 7.3.4).

At $\alpha = 21$ there is a region of two genes ($p = 0.0485$), *LY9* and *SH2D1B*, which increases to three genes with the addition of *CD244* at $\alpha = 24$ ($p = 0.0015$). These genes' rankings reflect a novel frameshift deletion in *LY9* in exome S0305 and a novel heterozygous missense SNV in *SH2D1B* in exome S0339, both of which are the highest ranked genes in the neighbourhoods of all three genes in the respective exomes. In terms of common function, a GO enrichment test finds *LY9* and *SH2D1B* to be annotated with "lymphocyte mediated immunity" (*adjP* = 0.0044), while all three genes are involved in "immune system process" (*adjP* = 0.0102).

At $\alpha = 61$ there is a region of four genes ($p = 0.0027$): *ATN1*, *SLIT1*, *SSPO* and *ZNF862*. In exome S0333 *SSPO* contains a heterozygous frameshift deletion that was only seen once in the in-house exome database, making it the best-ranked gene in the neighbourhoods of all four genes in this exome. In exome S0336 *ATN1* contains a novel heterozygous non-frameshift insertion, making it the best-ranked gene in the neighbourhoods of *SLIT1* and *ZNF862* for this exome (although *ATN1* itself has its rank adjusted due to a better-ranked neighbour, *GNAS*, which contains a novel heterozygous frameshift deletion). An enrichment test on these four genes finds no GO terms to be

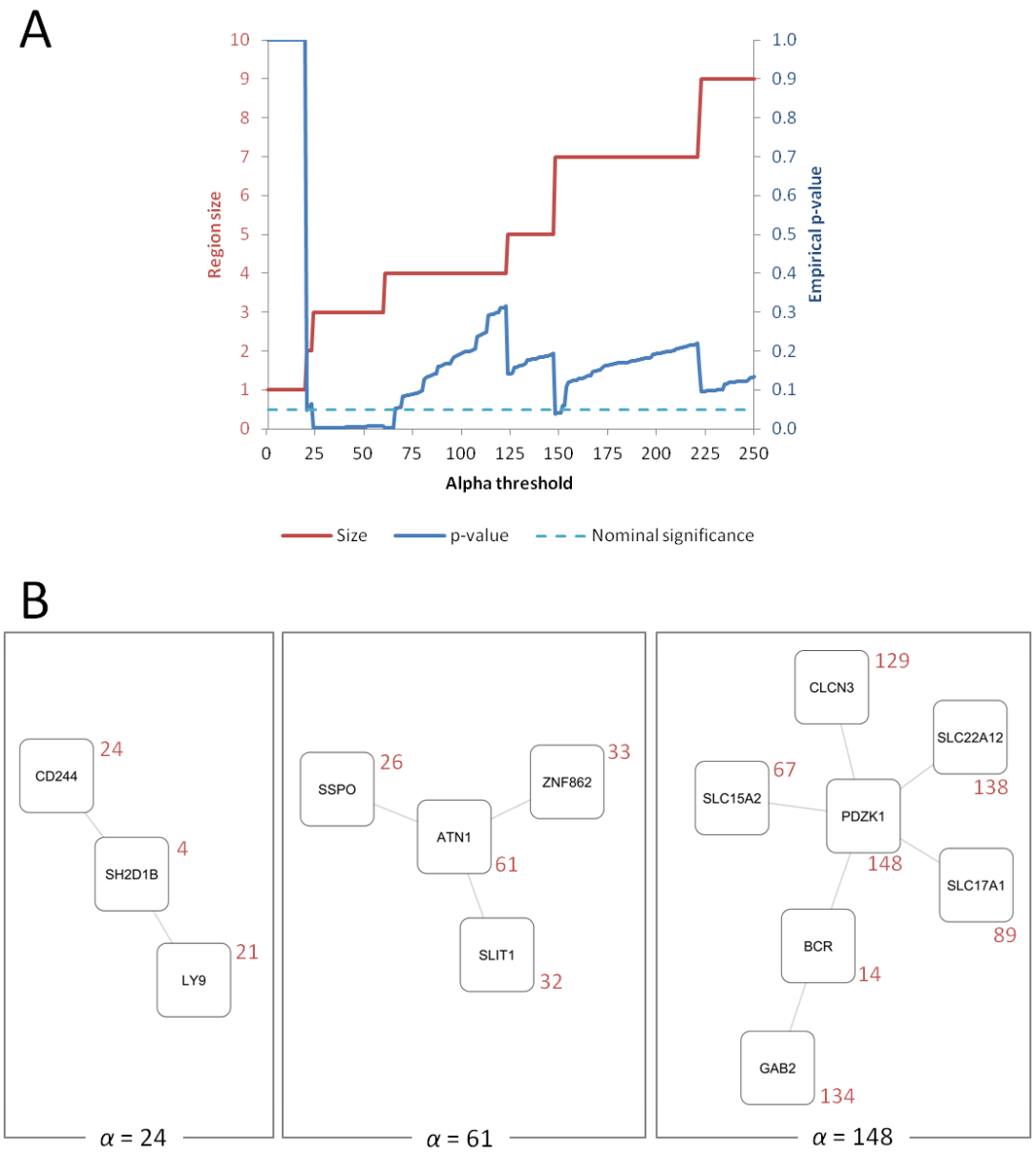


Figure 7.18 – RGA output based on PINA HetRank rankings for 13 unsolved AOS exomes
(A) For each alpha value, size of network region and empirical p-value are plotted. The nominal significance threshold is plotted at 0.05. This output is characteristic of RGA results: the region size is an increasing step function; at each alpha value where there is an increase in region size, the empirical p-value falls; between such alpha values the empirical p-value steadily increases. (B) Nominally significant regions found at alpha levels 24 (three genes, of which *SH2DB1* and *LY9* form the significant region of two genes at α = 21), 61 and 148. Final ranks are indicated in red.

significantly overrepresented; the most enriched annotation is "regulation of neuron differentiation" (*ATN1* and *SLIT1*; *adjP* = 0.1716).

At α = 148 there is a region of seven genes (p = 0.0387): *BCR*, *CLCN3*, *GAB2*, *PDZK1* and three solute carrier genes *SLC15A2*, *SLC17A1* and *SLC22A12*. In exome S0302 *BCR* harbours a low-frequency heterozygous missense SNV and is the best-ranked gene in the neighbourhoods of the three solute carrier genes; in S0306 *BCR* contains a heterozygous

frameshift insertion seen 28 times in the in-house exome database and is the best-ranked gene in the neighbourhood of *GAB2* (although the six other genes are all at least indirect neighbours of *CFTR* which contains a novel heterozygous frameshift deletion); the same frameshift insertion is found in S0333, making *BCR* the best-ranked gene in the neighbourhoods of *CLCN3*, *PDZK1* and the three solute carrier genes. Notably in both of the exomes S0304 and S0308, five of the genes in the region have their ranks adjusted due to variants in *SLC22A11*, which is not part of the region (it has an overall rank of 280 in the HetRank results). In terms of a shared function, all seven genes are annotated with the high-level GO term "transport" (*adjP* = 0.0014); we also see two genes involved in the "urate metabolic process" (*SLC17A1* and *SLC22A12*; *adjP* = 0.0013) and two more in "regulation of myeloid leukocyte mediated immunity" (*BCR* and *GAB2*; *adjP* = 0.0014).

These results demonstrate that from the HetRank results RGA can pick out network regions in which different genes boost each other's rankings because they contain variants in different exomes. However, it is also the case that genes just outside of a found region can be direct or indirect neighbours of most or all region genes. Therefore a severe variant in such a gene for one exome can boost the rankings of most or all region genes, without this neighbour being picked up by RGA. Arguably, then, RGA is only partially effective as a means of making sense of HetRank results.

Besides RGA, another approach to systematically analyse the genes ranked highest by HetRank is to look for shared functional annotation. However, a GO enrichment test of the top 20 genes (those in Table 7.10) found no significantly enriched GO biological process terms, the lowest adjusted p-value found being for "chloride transport" (genes *FXYD3*, ranked 12th, and *SLC26A8*, ranked 16th; *adjP* = 0.1386). If the test is instead performed using all genes ranked in the top 250 (consistent with the range of alpha values used for RGA), we find several significant terms. 15 genes are annotated with "regulation of membrane potential" (*adjP* = 0.0069), the highest-ranked such gene being *CAV3* (ranked 23rd). Three genes are annotated with "maintenance of cell polarity" (*ANK1*, ranked 82nd; *ATN1*, ranked 61st and which was picked up by RGA, and *DST*, ranked 167th; *adjP* = 0.0145). 31 genes are annotated with "cell adhesion" (*adjP* = 0.0145), the highest-ranked being *MUC4* (ranked 19th) but also including *LY9* and *SSPO* which were picked up (albeit in different regions) by RGA. Finally, four genes were annotated with "retinal ganglion cell axon guidance" (*adjP* = 0.0221), the highest-ranked being *SLIT1* which was also picked up by RGA. It is not immediately clear whether any of these biological process terms are relevant to AOS aetiology; whether or not the genes involved should be investigated further could be ascertained by a more detailed examination of the variants they harbour in the AOS exomes.

A final note on the HetRank results using PINA and the 13 unsolved AOS exomes regards the ranking of genes that are not in the network. The gene *EXO5* was ranked third overall by HetRank without any network-based rank adjustment (see Table 7.9). However, in the final network-adjusted results it is assigned a rank of 6,036. As would be expected, it has the highest rank among non-network genes. What this indicates is that network genes tend to receive a substantial boost to their ranks. Genes have a great advantage in HetRank by simply being part of the network, because the relatively large neighbourhood sizes give a high likelihood that a better-ranked neighbour will exist and a gene's rank will be adjusted upwards. In particular, this is compounded when the final rankings are calculated by combining adjusted ranks across all exomes. *EXO5* had pre-adjustment ranks of 17 in exome S0306, 19 in S0308, 18 in S0339 and 144 in the shared S0039/S0301 exome. After adjustment these drop to 77, 62, 75 and 2,151 respectively due to the effect of adjustments to network genes. While the first three of these ranks are still relatively high, there are nine other exomes in which *EXO5* did not start with such a high rank and where a network neighbourhood of several hundred genes would have offered a very good chance of an adjustment. Although the rank-adjustment process in HetRank does limit the advantage bestowed on genes with large neighbourhoods, the allowance is clearly not sufficient to allow non-network genes to be assigned a reasonable final rank. On the other hand it is not clear how this could be rectified in an unbiased way; after all, if variants in neighbouring genes in the network do constitute evidence for a shared role in AOS (as HetRank assumes), then genes in the network are indeed predisposed to have more evidence of this type. It is worth pointing out that if a non-network gene were ranked first in each exome before the ranks are adjusted (or indeed simply ranked ahead of the highest-ranked network gene in each exome), then it will still rank ahead of all network genes after the adjustment and hence in the final combined rankings. The paradox is that this situation would most likely arise for a disease where only a single gene is causal, so that there is no locus heterogeneity. In this case the gene would likely be identified using simple intersection filtering, and HetRank would not be required at all.

7.3.7.2 HetRank Analysis Using all 19 AOS Exomes

To judge how well HetRank picks out the known AOS genes, and to see how the results change from the previous section when additional exomes are added, we can examine the results of performing HetRank analysis on all 19 AOS exomes using the PINA network.

As before, it is instructive to start with the rankings obtained without any network-based adjustment. The top 20 genes ranked in this way are shown in Table 7.11, which also

Table 7.11 – Top 20 genes ranked by HetRank without network-based rank adjustment, based on all 19 AOS exomes
"Prev. rank" = previous rank assigned to the gene based on 13 unsolved AOS exomes only (i.e. by HetRank without network-based rank adjustment; the top 20 genes were summarised in Table 7.9)

Rank	Gene	In PINA?	Prev. rank
1	*SLC9B1*	✗	6
2	*AGAP6*	✗	12
3	*ADAMTS7*	✗	17
4	*VCX*	✗	8
5	*ABCA10*	✓	1
6	*EXO5*	✗	3
7	*FAM86B2*	✓	18
8	*PUSL1*	✓	42
9	*COTL1*	✓	7
10	*ARHGEF5*	✓	21

Rank	Gene	In PINA?	Prev. rank
11	*AQR*	✓	27
12	*PRAMEF1*	✗	5
13	*NINL*	✓	4
14	*ATN1*	✓	33
15	*RGPD3*	✓	10
16	*NOTCH1*	✓	1,843
17	*ZDHHC13*	✓	2
18	*ITPR3*	✓	126
19	*NBPF14*	✓	73
20	*SGK223*	✓	134

lists the ranks they were assigned by the same procedure using the 13 unsolved AOS exomes only (cf. Table 7.9).

There is substantial overlap between this top 20 and the top 20 genes based on the 13 unsolved cases (12 genes are common to both), as would be expected given the overlap in the underlying exome data. Of the top 20, 19 were previously ranked 134 or better, suggesting that the procedure used by HetRank to combine rankings across exomes is reasonably robust to the addition of new exomes.

As an example of the effect of this additional exome data, consider the top-ranked gene *SLC9B1*. This gene was ranked sixth when HetRank analysis was performed without a network adjustment for the 13 unsolved AOS exomes, due in large part to a heterozygous stopgain SNV that is found in exomes S0302 and S0333 (albeit observed 17 times in the in-house exome database). However, in the additional exomes *SLC9B1* has high ranks in S0311 (for which AOS is known to be caused by *NOTCH1*) and S0337 (AOS caused by *RBPJ*) due to the same variant. Leaving aside the fact that this specific variant would be ruled out by the previous observations in the in-house database, the assumption that only one variant per exome causes AOS already implies that *SLC9B1* cannot have an AOS role in the solved cases. So these rankings are impacted by non-causal variants in the solved cases.

The gene in the top 20 with the biggest jump in rank due to the addition of the six solved cases is *NOTCH1*, ranked 16th (and previously ranked 1,843). Encouragingly, this is caused primarily by the novel heterozygous frameshift insertion in exome S0311 and the novel heterozygous frameshift deletion in exome S0335 that do cause AOS (and which both cause *NOTCH1* to be ranked joint-top in their respective exomes).

The other known AOS genes do not rank in the top 20. *ARHGAP31* is assigned a rank of 839, which is considerably higher than its previous rank of 8,916 due to the novel heterozygous nonsense mutation that causes AOS in exome S0038 (*ARHGAP31* is ranked joint-top in this exome). *DOCK6* is ranked 520, which compares to 3,502 previously. This is due to a novel heterozygous missense SNV ranked 164 in exome S0332 (one of the two variants which together form a compound heterozygous mutation that causes AOS in this individual), and a novel heterozygous splice-site variant ranked 162 in exome S0334 (one of two variants causing AOS; note the other was not captured by whole exome sequencing). *RBPJ* is assigned a rank of only 4,138, which compares to 10,041 previously; the variant which causes AOS in S0337 is a novel heterozygous missense SNV which gives *RBPJ* a rank of 37 in this exome. Finally *EOGT*, which does not cause AOS in any of the six solved cases, has a rank of 10,575 due to various low-ranking variants in eight of the exomes (compared to a previous rank of 11,484 due to low-ranking variants in five of the unsolved cases).

Table 7.12 shows the results of performing HetRank analysis on all 19 AOS exomes using the PINA network. The "pre-network rank" column shows that the network-based adjustment causes considerable changes in the rankings, with many genes in the top 20 having had their rank adjusted by several thousand places (we previously saw the same thing in Table 7.10 based on only the 13 unsolved exomes).

The "previous rank" column allows comparison with the final rank (after network-based adjustment) that was assigned to each gene by HetRank using the 13 unsolved cases only (discussed at length in section 7.3.7.1 above). Six of the previous top 20 genes are still in the top 20 here (for example, *POLN*, which was the top-ranked gene based on the 13 unsolved exomes and is ranked sixth here based on all 19 AOS exomes). On the other hand, there are several genes now in the top 20 which had a rank above 1,000 previously. Again the gene with the biggest improvement in position is *NOTCH1*, ranked seventh (compared to 2,166 previously). This is primarily due to the AOS-causing variants that it contains in S0311 and S0335; it is also helped by an adjusted rank of 266 in exome S0038 due to a better-ranked neighbour (*NOTCH3*), and notably due to a relatively good adjusted rank of 342 in exome S0337 – which is due to the AOS-causing variant in its direct neighbour *RBPJ*.

This is a situation where HetRank works exactly as it is designed to do: the underlying functional relationship between *NOTCH1* and *RBPJ* is captured by an edge in PINA; consequently *NOTCH1*'s final rank, which is already high pre-adjustment because of the AOS-causing variants in S0311 and S0335, is boosted due to the AOS-causing variant that *RBPJ* harbours in S0337. This can be considered a proof-of-principle for HetRank: had

Table 7.12 – Top 20 genes ranked by HetRank using PINA network, based on all 19 AOS exomes
"Rank" = overall ranking combined across all 19 exomes after network-based adjustment; "# top 200 ranks" = number of the 19 exomes in which gene is ranked ≤200; "Prev. rank" = previous rank assigned to the gene based on 13 unsolved AOS exomes only (i.e. by HetRank using PINA network; the top 20 genes were summarised in Table 7.10); "Pre-network rank" = overall ranking combined across all 19 exomes before network-based adjustment (the top 20 such genes were shown in Table 7.11).

Rank	Gene	# top 200 ranks	Prev. rank	Pre-network rank	# top 200 ranks (pre-)
1	*CAPN11*	4	54	22	4
2	*MAML2*	2	668	364	1
3	*COL23A1*	2	684	3,276	0
4	*DBH*	3	8	853	0
5	*FCGR3A*	3	2	284	0
6	*PPP1R26*	3	560	727	1
7	*NOTCH1*	2	2,166	16	2
8	*POLN*	3	1	150	3
9	*MMP12*	1	90	908	0
10	*LFNG*	2	2,098	6,006	0
11	*MAML3*	2	1,151	3,117	1
12	*ABCA10*	5	5	5	5
13	*HGFAC*	2	6	1,672	0
14	*SH2D1B*	3	4	1,534	1
15	*GOLGA8A*	2	103	59	4
16	*ABCB5*	3	188	80	2
17	*GIF*	4	555	2,352	1
18	*HPCAL4*	2	323	5,109	0
19	*JAG2*	2	1,010	4,949	0
20	*MFNG*	2	1,490	2,959	2

we not already known any of the causal genes for AOS, an inspection of the top ten genes identified by HetRank would have put forward *NOTCH1* as a candidate.

To explore why other genes undergo such an improvement in rank compared to the ranks based on the 13 unsolved exomes, consider the next-best improvement. *LFNG* was ranked 2,098 previously, and only ranked 6,006 based on all 19 exomes before the network adjustment, yet has a final adjusted rank of 10. As might be expected, this is because *LFNG* is a direct neighbour of *NOTCH1* and its rank is boosted due to this association (in fact it only has two direct neighbours, so the adjustment weights it heavily toward *NOTCH1*'s scores in the relevant exomes). But since *LFNG* does not have a variant that confers it a high rank in any of the unsolved cases, it can be considered a false positive finding here.

We saw that *NOTCH1*'s rank was boosted due to the true causal variant in *RBPJ*. The converse also holds: *RBPJ*'s rank is boosted from 4,138 to 897 by the network-based adjustment, mainly due to the causal variants that *NOTCH1* contains in S0311 and S0335. However, the true causal variants in the three exomes for which AOS is caused by *RBPJ* and *NOTCH1* do not provide sufficient signal for HetRank to rank *RBPJ* anywhere near the top 20 genes, since the rankings are also based on 16 other exomes for which these are not the genes responsible for AOS. This means that if we did not already know that *RBPJ* was a causal gene for AOS, this analysis would not identify it.

(If HetRank analysis is performed using only the three exomes for which AOS is caused by *NOTCH1* or *RBPJ*, *NOTCH1* is ranked top and *RBPJ* is ranked 20th [full results not shown]. This suggests that *RBPJ* would be much more likely to be found by HetRank were the locus heterogeneity of AOS limited to these two genes.)

ARHGAP31 is assigned a rank of 272, up from 839 before the network adjustment: as well as the true causal variant it harbours in S0038, its indirect neighbour *ITSN2* is ranked joint-top for exome S0332 (actually known to have AOS due to a compound heterozygous mutation in *DOCK6*) and its direct neighbour *ITSN1* is ranked 30th for S0336 (an unsolved exome).

DOCK6 is ranked 10,216, despite a rank of 520 before the network adjustment. This comes about because *DOCK6* does not have any network neighbours that substantially boost its rank in any of the 19 exomes (in particular, while *DOCK6* is an indirect neighbour of *NOTCH1*, this link is via the exceptionally high-degree node *UBC* which implies that *DOCK6* has a very large indirect neighbourhood and hence the extent of any adjustment in the exomes for which *NOTCH1* causes AOS is very small). For exomes S0332 and S0334, in which *DOCK6* does contain AOS-causing variants, its ranks fall from 164 and 162 to 1,034 and 1,148 respectively, as a result of improvements to other genes' ranks elsewhere in the network.

Considering some of the very top-ranked genes, *CAPN11* is ranked first due to a heterozygous frameshift insertion it contains in the unsolved cases S0308 and S0336, and variants in its neighbours *QRSL1* and *NOTCH1* in the solved cases S0334 and S0335 respectively (with the latter being a true causal variant). This would make *CAPN11* a good candidate for follow-up study, except that the frameshift insertion has previously been seen in 20 other non-AOS exomes in the in-house exome database; this demonstrates that this variant does not cause AOS and thereby removes any direct evidence for *CAPN11*'s involvement in the disease process. *MAML2* is ranked second, mainly because it is a direct neighbour of *NOTCH1* and has a high adjusted rank in the two solved cases for which *NOTCH1* is causal (although *MAML2* itself contains a novel heterozygous non-frameshift

deletion in the unsolved case S0306). *COL23A1* is ranked third, due primarily to a variant in its indirect neighbour *ITSN2* in the solved exome S0332 (where *DOCK6* is causal) and the true causal variant in its indirect neighbour *NOTCH1* in exome S0335. To a lesser extent, variants in indirect neighbours in four other exomes also contribute to its high rank – although one of these variants is in a non-causal gene in the solved case S0337. Since *COL23A1*'s pre-network rank was not especially high there is little direct evidence that it plays a role in AOS.

We can again look for an underlying disease mechanism by employing RGA or by testing whether relevant functional annotation exists for highly-ranked genes.

RGA was carried out using all alpha thresholds in the range $1 \leq \alpha \leq 250$, and every threshold from 11 (region of four genes) up to 250 (region of 24 genes) gave highly significant results ($p < 0.0001$). Unlike what we saw based on the 13 unsolved exomes only (Figure 7.18 in the previous section), the significant regions found at increasing alpha thresholds form a nested series (see Figure 7.19). At lower alpha thresholds the regions form around *NOTCH1* (such as the four-gene region at $\alpha = 11$ comprising *LFNG*, *MAML2*, *MAML3* and *NOTCH1*), but as the alpha threshold increases the regions grow to incorporate a set of four genes previously identified by RGA based on HetRank in the 13 unsolved exomes (*ATN1*, *SLIT1*, *SSPO* and *ZNF862*; Figure 7.18b, $\alpha = 61$) and subsequently an extension which includes *LPA* (ranked 39th), which was highly ranked by HetRank using the 13 unsolved exomes (Table 7.10) and was previously a key gene in several of the optimal BioGranat-IG subnetworks (for example, see Figure 7.7).

To investigate whether this region is functionally relevant to AOS, a GO enrichment test was performed on all 24 genes. Unsurprisingly, "notch signalling pathway" is the most enriched term, covering nine genes in the region (*NOTCH1*, *NOTCH2* and direct interactors; $adjP = 6.91 \times 10^{-10}$), with four of them also involved in "notch receptor processing" ($adjP = 1.30 \times 10^{-5}$). The other significantly enriched terms are also related to notch-signalling, including "cardiac atrium morphogenesis" (*DLL4*, *NOTCH1* and *NOTCH2*; $adjP = 0.0007$) and "morphogenesis of an epithelial sheet" (*DLL4*, *MMP12*, *NOTCH1* and *NOTCH2*; $adjP = 4.76 \times 10^{-5}$), both of which are consistent with the cardiac defects seen in some cases of AOS (Snape et al. 2009). Other than this last term, which covers *MMP12*, the only significantly enriched biological process annotation was for *NOTCH1*, *NOTCH2* and direct neighbours, which does not suggest a close functional link with any of the more peripheral genes in the region. It is important to note the possibility that several of the genes in the region (particularly those connected to *NOTCH1*) have little direct evidence that they are involved in AOS and HetRank only ranks them highly due to variants in their neighbours (as we previously saw with the third-ranked gene *COL23A1*). However, given the

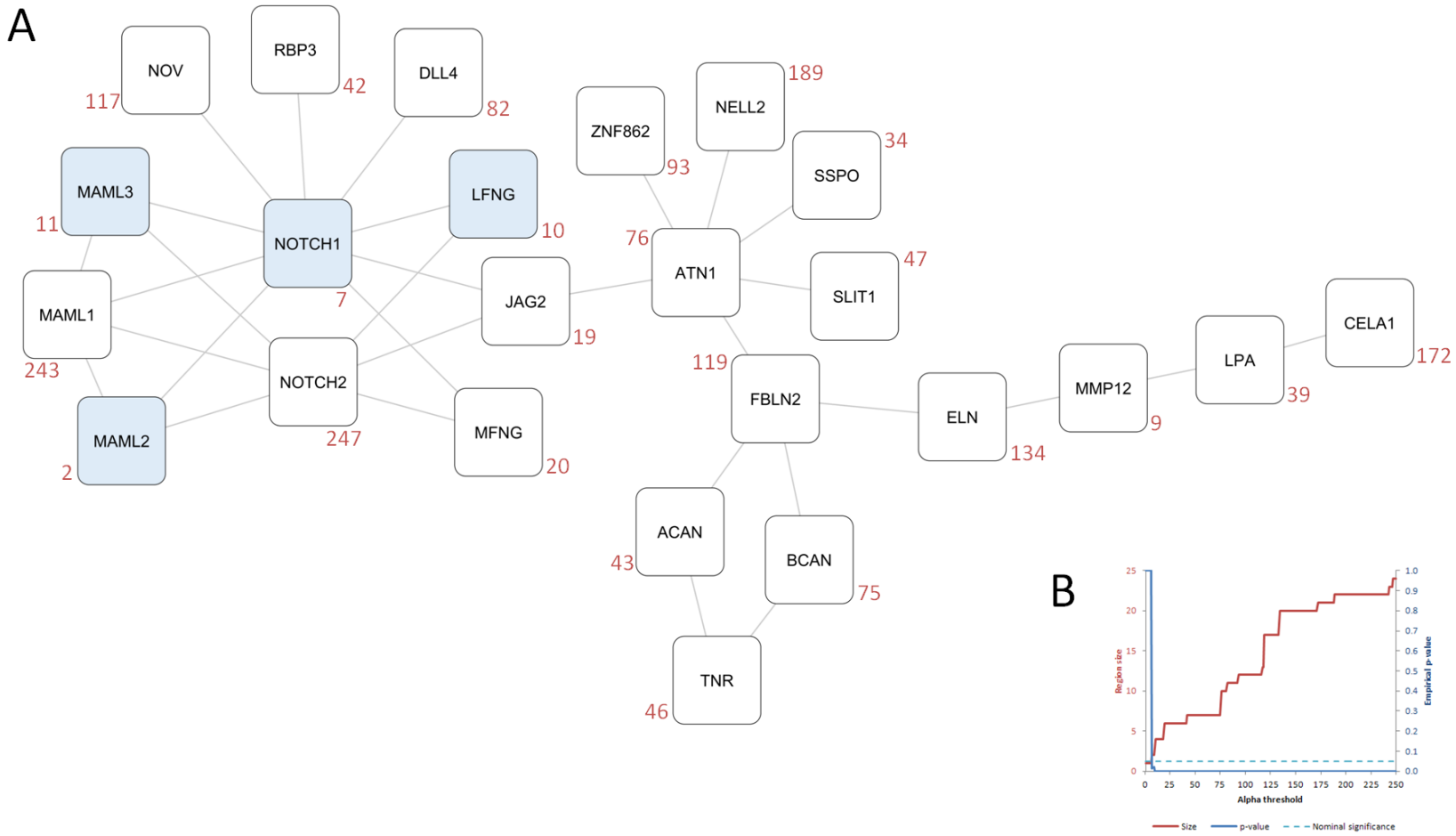


Figure 7.19 – RGA output based on PINA HetRank rankings for all 19 AOS exomes
(A) Blue nodes: highly significant region of four genes at $\alpha = 11$; at subsequent alpha thresholds this region is extended and remains highly significant right through to the maximum alpha value tested ($\alpha = 250$; all 24 genes). Final ranks are indicated in red. (B) Plot of region size (red) and empirical p-value (blue) by alpha value (same format as plot shown in Figure 7.18a) clearly shows that region size is nominally significant over almost the entire range.

demonstrated role that notch signalling genes play in AOS (Hassed et al. 2012; Stittrich et al. 2014) the genes in this region could benefit from further investigation.

Aside from RGA, the other way we can interpret HetRank results is to perform an enrichment test directly on the top-ranked genes. Based on the top 20 genes, "notch signalling pathway" (*JAG2*, *LFNG*, *MAML2*, *MAML3*, *MFNG* and *NOTCH1*; $adjP = 4.02\times10^{-6}$) and "morphogenesis of an epithelial sheet" (*MMP12* and *NOTCH1*; $adjP = 0.0452$) are the significantly enriched GO biological process terms, which is consistent with the RGA results. If we look at the top 250 genes, "notch signalling pathway" is still significantly enriched (11 genes; $adjP = 0.0029$), but so is "cell adhesion" (50 genes; $adjP = 1.32\times10^{-11}$) and its child term "homophilic cell adhesion" (13 genes; $adjP = 9.51\times10^{-5}$), "endocytosis" (19 genes; $adjP = 0.0014$), "cell morphogenesis involved in neuron differentiation" (24 genes; $adjP = 0.0032$), and "extracellular structure organisation" (13 genes; $adjP = 0.0041$). Depending on the plausibility that deficiencies in any of these functions could result in the phenotypes associated with AOS, these gene-sets could represent several different avenues for follow-up work.

7.3.8 HetRank Results: Other Networks

We now consider the results of carrying out HetRank analysis on the 13 unsolved AOS exomes, using the other four networks to adjust ranks. Table 7.13 gives the top 20 ranked genes for the higher confidence PINs, PINAmin2 and CPDBconf95, and for the COXPRES30 and Multinet networks, and compares the results to those obtained using PINA.

It is immediately clear that there is little overlap between the top-ranked genes using the different networks: no genes are ranked in the top 20 by both PINA and PINAmin2, with only two, two and three of the top 20 PINA-ranked genes being ranked in the top 20 using CPDBconf95, COXPRES30 and Multinet, respectively. This is illustrated further by Figure 7.20 which depicts the final ranks based on each alternative network of the 200 genes which were ranked highest using PINA. For all four of the alternative networks, many of the top 200 PINA-ranked genes are assigned ranks of 2,000 or more. We previously saw in section 7.3.7.1 that the network-based adjustment (using PINA in that case) had a volatile effect on the unadjusted rankings. The same is true for the other four networks (Table 7.13), but Figure 7.20 also tells us that the post-adjustment rankings for different networks are highly variable relative to each other. One trend that can be picked out from the figure is that genes which rank relatively highly across several of the networks (these are easier to pick out for the top 20 genes; e.g. *POLN*, *ABCA10*, *LPA*, *BCR* and *MST1*) tend to be those which were ranked relatively highly without any network adjustment (final column). This implies that

Table 7.13 – Top 20 genes ranked by HetRank using remaining four networks, based on 13 unsolved AOS exomes
"Pre-net." = overall ranking combined across all 13 exomes before network-based adjustment (top 20 such genes given in Table 7.9); "PINA rank" = corresponding rank when adjustment uses PINA network (top 20 such genes given in Table 7.10).

Rank	PINAmin2			CPDBconf95			COXPRES30			Multinet		
	Gene	**Pre-net.**	**PINA rank**	**Gene**	**Pre-net.**	**PINA rank**	**Gene**	**Pre-net.**	**PINA rank**	**Gene**	**Pre-net.**	**PINA rank**
1	*DNM1*	943	1,740	*NINL*	4	88	*NINL*	4	88	*ABCA10*	1	5
2	*GRAP2*	4,760	2,407	*MAP3K11*	46	248	*SP140L*	186	7,511	*NINL*	4	88
3	*LRP2*	81	222	*LRP2*	81	222	*ZDHHC13*	2	2,514	*POLN*	25	1
4	*CTTN*	2,419	4,574	*XRCC1*	533	3,790	*MST1*	22	17	*COTL1*	7	3,332
5	*ARHGAP32*	676	1,880	*DST*	457	167	*GJB2*	9	431	*GRIK1*	107	7,711
6	*ALK*	1,419	155	*CROCC*	859	6,050	*RTTN*	3,823	12,548	*PER2*	3,904	5,821
7	*SHB*	7,636	698	*APPL2*	430	286	*GOLGA8A*	43	103	*ARHGEF5*	21	2,154
8	*WASL*	2,403	1,732	*TEP1*	97	5,314	*WDR67*	88	7,385	*ZNF679*	616	40
9	*ITGB4*	166	714	*KRT83*	544	12,737	*AS3MT*	776	10,154	*GRIK2*	1,106	396
10	*PIK3AP1*	1,509	6,694	*FCGR3A*	224	2	*COTL1*	7	3,332	*SGK223*	134	6,648
11	*PTPN12*	304	1,145	*LAT2*	684	7,046	*LAPTM4B*	16	147	*CYP2F1*	700	13,048
12	*MAP4K3*	5,108	5,473	*KIDINS220*	645	5,861	*ANGPT2*	89	787	*HSPBP1*	93	1,338
13	*EPB41L2*	375	1,585	*CD3G*	756	102	*LPA*	98	7	*TEP1*	97	5,314
14	*SYNJ1*	1,772	866	*MUTYH*	2,814	9,562	*LRP2*	81	222	*DTX2*	79	4,598
15	*FANCD2*	1,756	2,015	*ALS2*	4,572	12,431	*PTPRG*	49	1,072	*TBXA2R*	1,345	3,731
16	*MLLT4*	3,467	2,999	*PKD1*	451	7,317	*GNRH2*	14	8,329	*ZNF358*	124	8,656
17	*CD22*	2,826	3,571	*BCR*	75	14	*PNLIPRP3*	28	9,324	*MST1*	22	17
18	*GAB2*	3,859	134	*AQR*	27	4,834	*ZNF358*	124	8,656	*CHST15*	140	5,453
19	*SORBS1*	5,358	6,022	*ECH1*	2,416	8,871	*ARHGEF5*	21	2,154	*ADAM22*	753	471
20	*MAPK8IP1*	5,421	292	*GJB1*	15,863	1,445	*PUSL1*	42	6,451	*ITGA1*	238	39

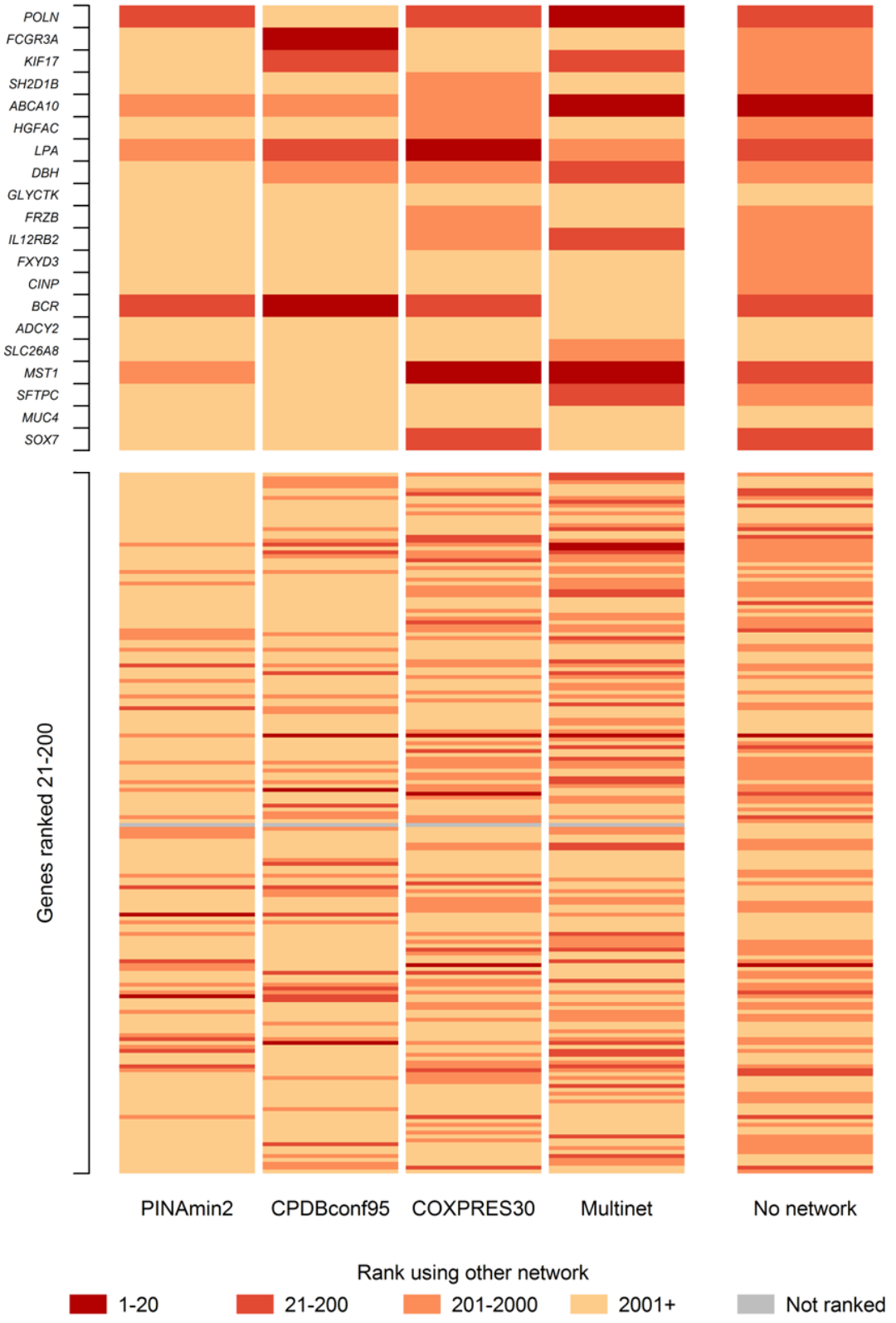


Figure 7.20 – Heatmap depicting rank using other networks of top 200 PINA-ranked genes, based on 13 unsolved AOS exomes
Y-axis = genes which are ranked 1-200 by HetRank using PINA for network-based adjustment; the top 20 genes (as in Table 7.10) are labelled individually; x-axis = other networks; colours represent rank of gene if other networks used for HetRank adjustment. The last column gives the corresponding rank if no network adjustment is performed.

such genes are consistently ranked highly due to the direct evidence for causality conferred by the variants they contain, and not due to indirect evidence inferred from network neighbours.

It is perhaps most surprising that the PINA results are not more consistent with those from PINAmin2, which is a higher-confidence subnetwork of PINA. This can be understood by looking more closely at the relative rankings. In Figure 7.21a the rank assigned to each gene when adjustments are based on PINAmin2 is compared to the rank assigned using PINA-based adjustments. As the discussion in section 7.3.7.1 would suggest, genes that are not in the PINAmin2 network tend to be assigned a low rank when HetRank uses PINAmin2, and so these genes (plotted in grey) generally undergo a fall relative to their PINA-based rank. Genes which are direct neighbours of the exceptionally high-degree node *UBC* in both networks (plotted in red) will have very large indirect neighbourhoods, and because the network-based adjustment takes neighbourhood size into account it will tend to have limited effect in both networks. On average, these genes experience an improvement in rank largely because of the number of genes that are not in PINAmin2 (and whose rank falls). Genes which are direct neighbours of *UBC* in PINA but not in PINAmin2 (plotted in green) tend to undergo a substantial improvement in their rank when PINAmin2 is used for adjustment, because their neighbourhood sizes are much smaller and therefore more weight can be given to neighbours which rank highly. Finally, for genes which are in both networks but not direct neighbours of *UBC* in either (plotted in blue), the changes to their network neighbourhoods may be smaller but they still have complex effects that cause dramatic changes in rank. This can be seen in Figure 7.21b: for each of the top 20 PINA-ranked and PINAmin2-ranked genes, this figure shows the number of individual exomes for which the same neighbour is used to adjust the rank (i.e. has the best unadjusted rank) in both networks. In a large majority of cases genes have their ranks adjusted due to different neighbours in the two networks.

To establish whether the HetRank results based on the other networks suggest any plausible new AOS genes, we can look at genes which are ranked highly using multiple networks, and we can look at RGA and functional enrichment test results for each network.

Three genes are ranked in the top 20 using three different networks. *LRP2* (PINAmin2, CPDBconf95 and COXPRES30) has been previously identified as a key gene in the BioGranat-IG results in PINA_d50 (see section 7.3.4) and *NINL* (CPDBconf95, COXPRES30 and Multinet) was singled out in section 7.3.2 due to the high number of variants it contains after filtering. *MST1* (PINA, COXPRES30 and Multinet) has not previously been discussed; it has a relatively high rank of 22 without any network adjustment due to a novel heterozygous nonsense SNV in exome S0304, but in every other

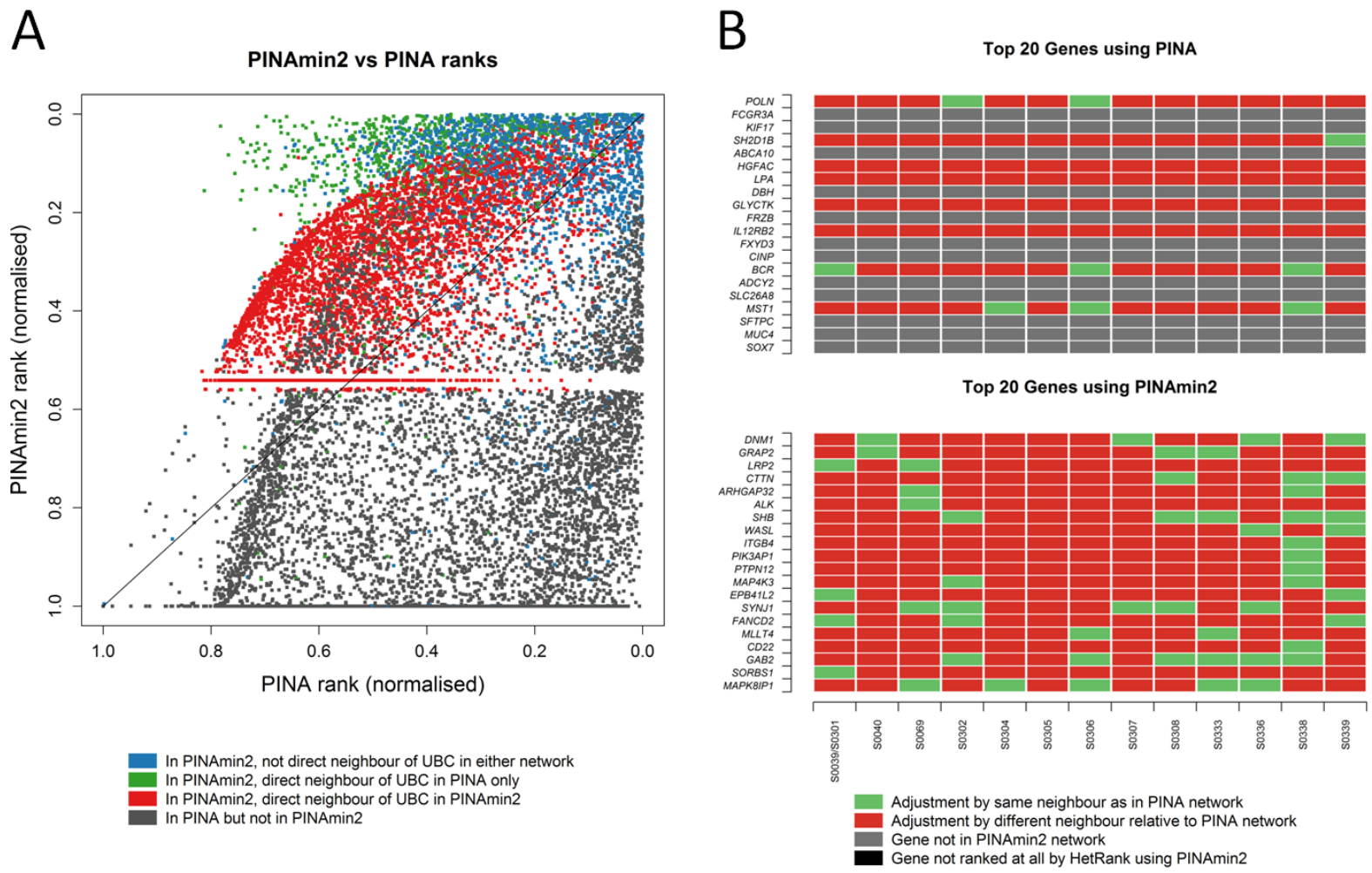


Figure 7.21 – Comparison of HetRank results based on PINA and PINAmin2, for 13 unsolved AOS exomes
See next page for full figure legend.

Figure 7.21 – Comparison of HetRank results based on PINA and PINAmin2, for 13 unsolved AOS exomes (previous page)
(A) For each gene in PINA, the final rank assigned by HetRank using PINAmin2 is plotted against the final rank using PINA. Ranks are normalised to range (0,1] and best-ranked genes are at top/right of plot. *UBC* has exceptionally high degree in both networks so to aid interpretation points are coloured according to whether they are direct neighbours of this node in both networks, in PINA only or in neither network. The red horizontal line at 0.5412 represents 550 genes that do not contain any variants themselves but are direct neighbours of *UBC* in PINAmin2 with no other direct neighbours (thus all receive an equal rank). The grey horizontal line at 1.0 represents 1,925 genes (13.4% of genes in PINA) that do not contain any variants and are not in PINAmin2 (thus not assigned a rank by HetRank using PINAmin2). (B) Agreement between PINA and PINAmin2 of neighbours used for adjustment, by individual exome. For each gene and exome, the colour indicates whether the gene's final rank in that exome was adjusted with respect to the rank of the same neighbouring gene in PINA and in PINAmin2.

exome is adjusted due to different neighbours for each exome (therefore does not point towards a consistent disease mechanism underlying AOS).

Nine genes are ranked in the top 20 using two different networks. For *ABCA10*, *BCR*, *FCGR3A*, *LPA* and *POLN*, one of these networks was PINA and these will not be discussed further. *ARHGEF5* (COXPRES30 and Multinet) stands out because it encodes a protein that regulates Rho GTPases (as does the AOS-causing gene *ARHGAP31*). It has a relatively high rank of 21 without any network adjustment, due to a heterozygous nonsense SNV in exome S0304 (seen 28 times in the in-house exome database) and a heterozygous frameshift insertion in exome S0333 (seen twice in the in-house exome database) – these previous observations would suggest these variants do not cause AOS. The other genes with top 20 rankings in two networks are *COTL1* (COXPRES30 and Multinet; pre-network rank of 7 and adjusted using different neighbours in every exome), *TEP1* (CPDBconf95 and Multinet; pre-network rank of 97 and adjusted using the same neighbour in five of the 13 exomes) and *ZNF358* (COXPRES30 and Multinet; pre-network rank of 124 and adjusted using the same neighbour in one of the 13 exomes).

Considering the results from each network in turn, the top 20 genes using PINAmin2 contain two that we previously saw in the optimal BioGranat-IG results for PINA_d50 and PINAmin2_d50 (see e.g. Figure 7.13): *LRP2* and *MAPK8IP1*. Both of these genes are highly ranked due to a number of network adjustments, particularly due to variants in *MAP3K11* (another gene identified by BioGranat-IG but ranked 981 by HetRank) for exomes S0069 and S0304 and in *MAPK9* for exome S0333, and due to the novel heterozygous splice-site variant that *LRP2* harbours in exome S0336. We also previously saw the top-ranked gene *DNM1* picked out by the simple neighbourhood search around the known AOS gene *ARHGAP31* in PINAmin2 (it is an indirect neighbour).

A GO enrichment test based on the genes ranked in the top 20 using PINAmin2 identifies one significant functional annotation: "receptor-mediated endocytosis" (*DNM1*,

SYNJ1, *WASL* and *LRP2*; *adjP* = 0.0056). When the top 250 genes are tested for functional enrichment, the biological process terms identified are highly significant, including "cellular response to stimulus" (162 genes; $adjP = 2.68\times10^{-22}$), "cellular communication" (150 genes; $adjP = 5.99\times10^{-19}$) and a child term of both, "signal transduction" (138 genes; $adjP = 1.05\times10^{-18}$). The fact that these terms are highly enriched could indicate that the network contains a region enriched for genes involved in signalling, and that plausible AOS-causing variants in some of these genes cause many of them to be highly ranked by HetRank. Performing RGA using the HetRank results enables this notion to be explored further.

Testing all alpha in the range $1 \leq \alpha \leq 250$, RGA finds several highly significant regions ($p < 0.0001$); since we seek a compact functional pathway underlying AOS we consider the smallest of these, which contains 14 genes and is presented in Figure 7.22. The gene that particularly stands out in this region is *ARHGAP32* because it encodes a GTPase-activating protein which acts on Cdc42 and Rac1, among other Rho GTPases (as does the AOS-causing gene *ARHGAP31*). However, *ARHGAP32* does not contain variants in any exomes that offer strong direct evidence for AOS causality; its high rank is due to variants in neighbouring genes, in particular its indirect neighbours *MAP3K11* (in exome S0069) and *BCR* (in S0306). An enrichment test based on the 14 genes in the region finds several equally significant GO biological process terms: "regulation of response to stimulus" (ten genes), "positive regulation of cellular component organisation" (six genes) and its child term "positive regulation of cell projection organisation" (*ABL2*, *MET*, *RET* and *WASL*) all have a multiple-test-adjusted p-value of 0.0036. Notably six of the genes are significantly annotated as "transmembrane receptor protein tyrosine kinase signalling pathway" (*CRK*, *EPS15*, *GAB1*, *GAB2*, *LCP2* and *RET*; *adjP* = 0.0044).

(If we consider the largest region found, of 71 genes at $\alpha = 247$ [$p = 0.0001$], a GO test shows these genes to be enriched for "signal transduction" [55 genes; $adjP = 2.39\times10^{-12}$], specifically its child term "transmembrane receptor protein tyrosine kinase signalling pathway" [29 genes; $adjP = 1.85\times10^{-14}$], as well as "phosphorylation" [36 genes; $adjP = 1.69\times10^{-12}$]. This supports the idea raised above that PINAmin2 contains a region enriched for functionally related genes with mutually supportive evidence for a role in AOS. It could therefore be beneficial to study these genes further.)

Among the top 20 genes using CPDBconf95, several have been previously identified by means other than HetRank: *NINL* (ranked top) contains a high number of post-filtering variants (see section 7.3.2); *MAP3K11* (ranked second) was identified by the simple neighbourhood search as a neighbour of *ARHGAP31* in Multinet (section 7.3.1), and subsequently found in BioGranat-IG results in PINA and PINAmin2 (sections 7.3.4 and 7.3.5), and *LRP2* (ranked third) was also a key gene in the BioGranat-IG results.

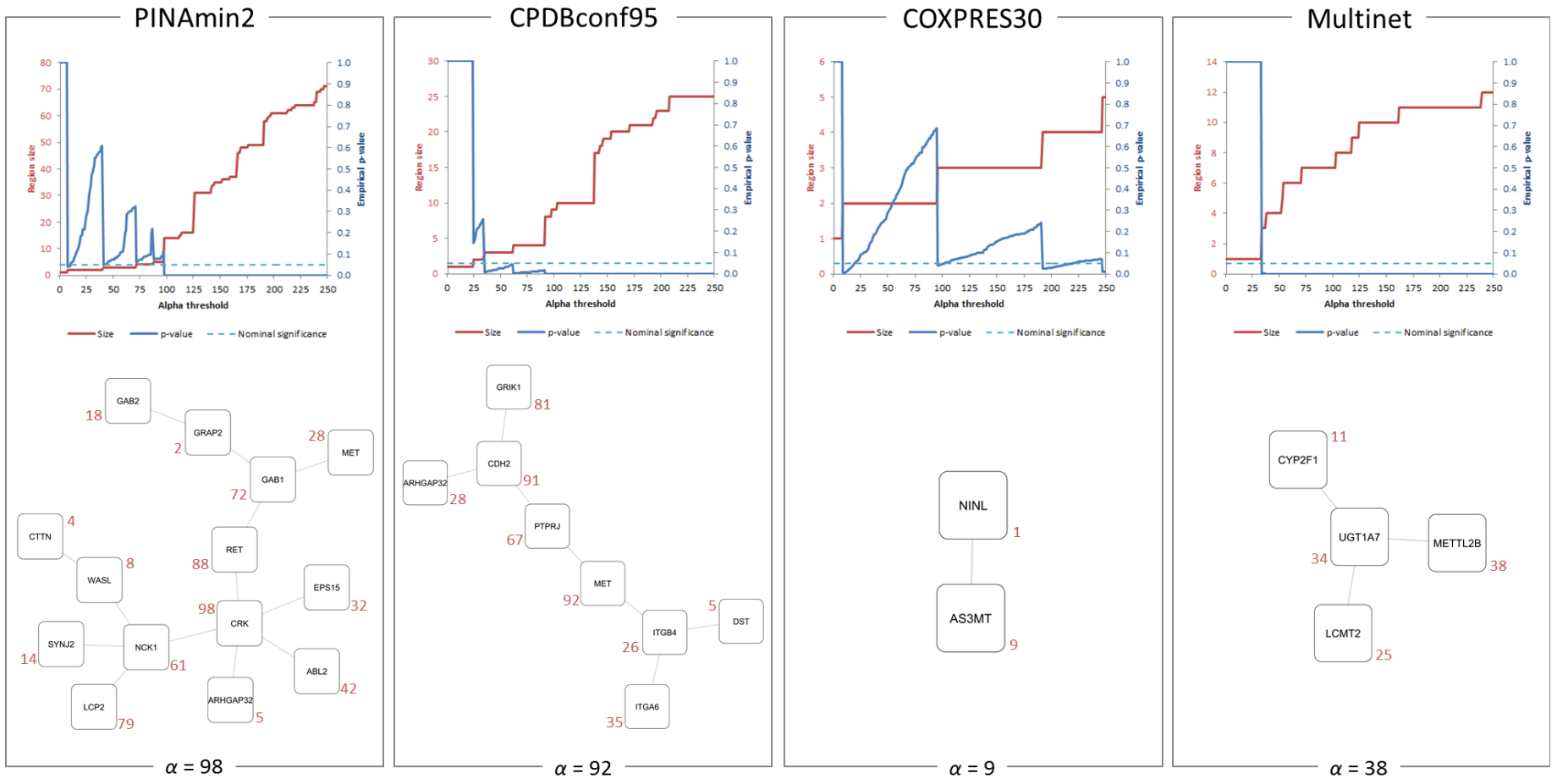


Figure 7.22 – RGA output based on HetRank rankings for 13 unsolved AOS exomes using other networks
For each network: top = plot of region size (red) and empirical p-value (blue) by alpha value, with same format as plot shown in Figure 7.18a; bottom = most significant network region found. Where multiple alpha thresholds result in equally significant regions (i.e. having the same empirical p-value), the smaller threshold is used (since we seek a compact network region). Final ranks are indicated in red.

A GO enrichment test using the top 20 genes finds no significantly enriched biological process annotation (data not shown). Based on the top 250 genes we find several equally significant terms with a multiple-test-adjusted p-value of 1.14×10^{-13}: "cell junction organisation" (27 genes), "cell adhesion" (57 genes) and "cellular component movement" (64 genes).

Testing all alpha in the range $1 \leq \alpha \leq 250$, RGA finds several highly significant regions ($p < 0.0001$). The smallest, a region of eight genes, occurs at $\alpha = 92$ (see Figure 7.22). Interestingly, two of the genes were also in the smallest RGA region based on the PINAmin2 results. As discussed above, *ARHGAP32* shows little direct evidence for AOS involvement; the other gene is *MET*, which is also assigned its relatively high rank of 92 based on variants in a number of neighbouring genes. A GO enrichment test of all eight genes finds that five of them (*CDH2*, *DST*, *ITGA6*, *ITGB4* and *PTPRJ*) are involved in "cell junction assembly" ($adjP = 3.03\times10^{-5}$), six of them (these five plus *MET*) are involved in "cell motility" ($adjP = 0.0007$), and three (*DST*, *ITGA6* and *ITGB4*) are part of the "integrin-mediated signalling pathway".

Again using COXPRES30, some of the genes in the top 20 have previously been identified by means other than HetRank. *NINL* is again ranked first; *ZDHHC13* (ranked third) was in Table 7.5 due to the high number of post-filtering variants it contains; *LPA* (13th) and *LRP2* (14th) were key genes in the optimal subnetworks identified by BioGranat-IG in PINA_d50 (section 7.3.4), and *PTPRG* (15th) was picked out in a region of near-optimal triplets that BioGranat-IG identified in PINA (Figure 7.11).

A GO enrichment test using the top 20 genes finds no significantly enriched biological process annotation (data not shown). Based on the top 250 genes we find one significantly enriched term: "homophilic cell adhesion" (11 genes; $adjP = 0.0095$).

When RGA is employed to test thresholds $1 \leq \alpha \leq 250$, the most significant region is found at $\alpha = 9$ ($p = 0.0053$). It comprises two genes, *NINL* and *AS3MT* (see Figure 7.22). The key drivers of *AS3MT*'s high rank appear not to be variants in *NINL* but variants in various other neighbours, although *AS3MT* itself harbours a novel splice-site variant in exome S0336. *AS3MT* encodes an arsenic methyltransferase, and it is not clear how this function could relate to that of *NINL*, which is involved in microtubule organisation in interphase cells. This illustrates a key problem with the COXPRES30 network: the edges are difficult to interpret. By querying the COXPRESdb database for *NINL* and *AS3MT* we see that they have a mutual rank of 7.3, making *AS3MT* the top co-expressed gene of *NINL* and *NINL* the third-most similarly expressed gene of *AS3MT*. This occurs because *NINL* and *AS3MT* have correlated expression across a range of gene expression samples. However, this

cannot tell us directly whether the genes encode proteins involved in some common process. Based on a GO enrichment test, both genes are involved in "cellular process" (*adjP* = 1.0).

Among the top 20 genes using Multinet, the only gene that has previously been picked out by other methods is *NINL* (ranked second).

In a GO enrichment test for the genes ranked in the top 20, *GRIK1* and *GRIK2* are both annotated with "regulation of short-term neuronal synaptic plasticity" and "negative regulation of synaptic transmission, glutamatergic" (*adjP* = 0.0055). Using the top 250 genes, the most significantly enriched terms are "regulation of excitatory postsynaptic membrane potential" (seven genes, including *GRIK1* and *GRIK2*; *adjP* = 0.0014), plus "cell adhesion" (34 genes) and its child term "homophilic cell adhesion" (12 genes; both with *adjP* = 0.0014).

Testing all alpha in the range $1 \leq \alpha \leq 250$, RGA finds several highly significant regions ($p < 0.0001$). The smallest, a region of four genes, occurs at $\alpha = 38$ (see Figure 7.22). The four genes are *CYP2F1*, *LCMT2*, *METTL2B* and *UGT1A7*, none of which we have previously identified by other methods. A GO enrichment test on the four genes suggests that *CYP2F1* and *UGT1A7* are involved in the "xenobiotic metabolic process" (*adjP* = 0.0018). The direct evidence linking these four genes to AOS is a novel heterozygous frameshift deletion that *LCMT2* harbours in exome S0069 and a novel heterozygous frameshift deletion that *UGT1A7* contains in exome S0305 (both of which are ranked joint-top in their respective exomes before the network-based adjustment). However, the functional annotation suggested by the enrichment test is not strongly supportive of a role in AOS.

As we have seen, the HetRank results from different networks support each other – and the results previously identified by BioGranat-IG – in a few instances, but often many new candidate genes are suggested. This emphasises the importance of a careful consideration of the evidence for a gene's involvement in the disease, which includes both the direct evidence (the nature of the variants identified by whole exome sequencing) and the indirect evidence from the network (the nature of the connections between genes and the relevance of any shared function that they undertake).

7.3.9 HetRank Results: Alternative Input Data and Parameters

The final results section looks at how the HetRank results change when either the input data or HetRank parameters (i.e. the weights assigned to different ranking factors) are modified. All results are obtained by performing HetRank analysis for the 13 unsolved AOS using the PINA network, and will be compared to the main results reported in section 7.3.7.1.

Table 7.14 – Top 20 genes ranked by HetRank before and after network adjustment using PINA, based on 13 unsolved AOS exomes with common variants removed
"All var. rank" = corresponding rank assigned to gene when HetRank analysis used the full exome data (top 20 ranked genes without network adjustment are in Table 7.9; top 20 genes after network adjustment are in Table 7.10); "Pre-net." = rank assigned to gene without network adjustment.

Without network adjustment			After network adjustment			
Rank	**Gene**	**All var. rank**	**Rank**	**Gene**	**Pre-net.**	**All var. rank**
1	*VCX*	8	1	*HGFAC*	864	6
2	*RGPD3*	10	2	*CD244*	5,326	24
3	*NINL*	4	3	*FXYD3*	1,413	12
4	*GJB2*	9	4	*POLN*	117	1
5	*ZDHHC13*	2	5	*GALNT7*	7,896	475
6	*MAPK6*	148	6	*ZNF440*	5,709	661
7	*TWISTNB*	64	7	*ZNF600*	3,910	141
8	*MST1*	22	8	*LIG4*	2,436	46
9	*GOLGA8A*	43	9.5	*KDSR*	13,558.5	453
10	*CNTNAP3B*	19	9.5	*FBXL16*	13,558.5	388
11	*MAP3K11*	46	11	*CDH13*	13,558.5	251
12	*BMP5*	56	12	*SIM1*	2,261	214
13	*MTX1*	165	13	*CA1*	13,558.5	922
14	*NBPF6*	20	14	*LHB*	6,487	53
15	*NBPF15,NBPF16*	15	15.5	*DLK2*	13,558.5	1,039
16	*GNRH2*	14	15.5	*EGFL9*	13,558.5	1,319
17	*SOX7*	72	17	*AFTPH*	9,735	372
18	*COTL1*	7	18	*ADCY2*	9,692	15
19	*ARHGEF5*	21	19	*FCGR1A*	4,167	1,318
20	*CCDC144NL*	11	20	*FCGR3A*	466	2

7.3.9.1 Alternative Input Data

AOS is almost certainly not caused by common genetic variants, due to its low prevalence, severe clinical phenotype and the monogenic causal mechanism previously identified in the known AOS genes. Therefore, it could be suggested that applying a weak filter to the whole exome sequencing data before it is analysed with HetRank might lead to more valid results by reducing the level of "noise" in the network-based rank adjustment step (meaning that genes will not have their ranks adjusted due to common variants in neighbours). In order for the HetRank tool design to remain valid, the input data must retain sufficient numbers of variants in case and control exomes (particularly in controls due to the nature of the control-derived ranking factor) to give a meaningful ranking of genes.

To this end, HetRank analysis was repeated with all common variants (defined as having either 1000 Genomes Project or EVS alternative allele frequency of ≥ 0.05) removed from the 13 unsolved AOS exomes and from all 336 non-AOS control exomes. The genes ranked in the top 20, both with and without the network-based rank adjustment, are presented in Table 7.14.

The top 20 genes before adjustment show a substantial overlap of 11 genes with the corresponding top 20 obtained when common variants were included (Table 7.9). This should be expected since high-ranking genes under both sets of input data (with and without common variants) should be those in which several different exomes contain novel or relatively rare nonsense, frameshift or missense variants – which would be present in both sets of data.

The top ranked gene is *VCX*, which was ranked eighth when the input data contained common variants. It contains variants in eight exomes (compared to all 13 including common variants) and its rank improves in seven of them, resulting in top-200 rankings in five exomes (compared to three previously). To understand why its rankings change we can consider one of the exomes in detail; in S0069 *VCX* is ranked 35.5 (compared to 78.5 previously). Since *VCX*'s rank improves, other genes' ranks must get worse – the biggest drop being for *MAN2A2* (from 10 based on all variants to 383 when common variants are excluded). *VCX* contains a heterozygous missense SNV annotated only with one previous observation in the in-house exome database, while *MAN2A2* contains a heterozygous nonsense SNV which is not in the in-house exome database but which has alternative allele frequencies of 0.0016 according to 1000 Genomes Project annotation and 0.0019 according to EVS. Since neither of these variants would class as common, they are present in both sets of data. The main reason that *MAN2A2*'s rank falls appears to be the ranking factor derived from control exomes: based on all variants, there are no non-AOS control exomes with a better (lower) ranking-score than *MAN2A2* has in exome S0069 but when common variants are excluded there are 11 non-AOS control exomes with a lower ranking-score (and given the high weight attached to the control-derived ranking factor, this causes *MAN2A2*'s rank to worsen). For *VCX* there are no such non-AOS control exomes for either set of data. This serves to illustrate the complex way in which variants in the non-AOS control exomes contribute to the final HetRank rankings.

Note that it would be a mistake to assume that common variants always result in a very poor ranking. For example, in exome S0336 the gene *OR4D10* has a relatively high non-network-adjusted rank of 111 when all variant types are included (due to a heterozygous nonsense SNV that has an EVS alternative allele frequency of 0.0763) but is unranked when common variants are excluded.

Considering now the final rankings after the network-based adjustment, a number of immediate observations can be made based on the top 20 genes in Table 7.14. Firstly, six of the genes have a non-network-adjusted rank of 13,558.5. These are all genes which contain no variants in any of the 13 unsolved AOS exomes when common variants are excluded (and HetRank therefore assigns them a "joint-last" pre-adjustment rank with the other 7,242 genes in PINA which do not harbour a variant). Their high final ranks are therefore entirely due to indirect evidence for AOS causality contributed by network neighbours. Secondly, for the first time among HetRank analyses we see genes in the top 20 having equal rank: for example, *KDSR* and *FBXL16* are both ranked 9.5. Neither of these genes contains a variant in any of the case exomes, and both have their ranks adjusted due to the same network neighbours in each exome. Further, they are both only connected to *MAPK6* in PINA which means they have identical direct and indirect neighbourhood sizes (meaning that their ranks are adjusted using the same weighting in each exome). The genes ranked 11th and 13th (*CDH13* and *CA1*) also have their ranks adjusted with reference to the same neighbours as these genes in every exome, but their final ranks are slightly lower due to having slightly larger neighbourhoods. We are seeing genes with equal final rank because, compared to when common variants were included, a greater number of genes have their ranks assigned based solely on indirect evidence from network neighbours, and therefore HetRank is less able to discriminate between genes. (Across the 13 unsolved AOS exomes there are on average 9,212.4 genes per exome that contain variants when common variants are included, but only 2,117.1 per exome when they are excluded.)

There are only five genes in common between the top 20 in Table 7.14 and the corresponding top 20 obtained when common variants were included (Table 7.10). This variability reflects the changes in the pre-adjustment ranks of individual genes as well as their network neighbours. The most notable change in rank is *EGFL9*, which had a final PINA-adjusted rank of 1,319 based on all variant types but has a rank of 15.5 when common variants are excluded. *EGFL9* itself contains no variants in any of the exomes and its high final rank is mainly due to relatively good adjusted ranks (between 160.5 and 292.5) in four of the exomes. When common variants are excluded, *EGFL9* undoubtedly benefits from the reduction in the number of genes containing variants, based on the following logic. If *EGFL9* contains no variants it will be ranked joint-last (pre-adjustment) in every exome. In any given exome more genes will contain no variants and hence be ranked joint-last when common variants are excluded, and because tied genes are given an average rank this means *EGFL9*'s joint-last rank is relatively better when they are excluded than when they are kept. Since a gene's adjusted rank in each exome is based not only on the unadjusted rank of its best-ranked neighbour but on its own unadjusted rank, this should make *EGFL9*'s network-

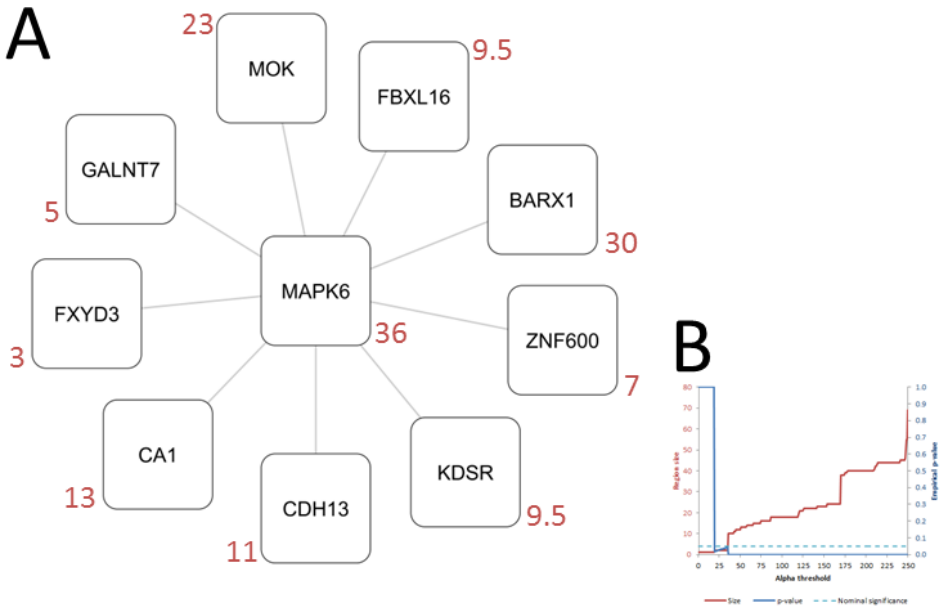


Figure 7.23 – RGA output based on PINA HetRank rankings for 13 unsolved AOS exomes with common variants removed
(A) RGA region found at $\alpha = 36$. Final ranks are indicated in red. (B) Plot of region size (red) and empirical p-value (blue) by alpha value (same format as plot shown in Figure 7.18a) clearly shows that region size is nominally significant over almost the entire range.

adjusted rank better in each exome and hence in the final combined prioritisation. This is arguably a sensible outcome, since there is little reason to believe that a gene containing a very common variant is more relevant to AOS than one containing no variant. (Note also that differences in the pre-adjustment ranks of neighbourhood genes resulting from inclusion or exclusion of common variants will also have an effect.)

To establish whether the final rankings suggest a different mechanism for AOS than has previously been suggested, we can look for enriched functional annotation in high-ranking genes, and employ RGA.

A GO enrichment test based on the top 20 genes finds no significantly enriched biological process terms (data not shown). Using the top 250 genes, the most significantly enriched terms include "regulation of membrane potential" (18 genes; $adjP = 7.79\times10^{-6}$), "regulation of synaptic transmission" (14 genes; $adjP = 0.0002$) and "glutamate receptor signalling pathway" (seven genes; $adjP = 0.0007$).

Testing all alpha in the range $1 \leq \alpha \leq 250$, RGA finds several highly significant regions ($p < 0.0001$). The smallest is at $\alpha = 36$ and comprises ten genes (see Figure 7.23). The region is based around *MAPK6*, which was ranked sixth before the network-based adjustment, mainly due to a heterozygous missense variant in exome S0069 that has not been previously seen in the in-house exome database or in 1000 Genomes Project or EVS data (although not technically novel because it has a dbSNP identifier), a novel heterozygous

splice-site variant in S0338 and a novel heterozygous SNV in S0339 which is synonymous but nevertheless relatively highly ranked (*MAPK6* has a pre-adjustment rank of 106 in that exome). These variants have a large influence on the high final ranks of the other nine genes in the region after adjustment. However, a GO enrichment test on these ten genes does not find any significant common functional annotation (data not shown)

7.3.9.2 Alternative Parameters

In section 7.3.7.1 we found that *ABCA10* was ranked first without the network-based adjustment, and fifth after adjustment, despite the fact that the heterozygous frameshift deletions it contains in four exomes had been seen 15-24 times in the in-house exome database. This suggests that the weight parameter used for the ranking factor "number of observations in heterozygous form in the in-house exome database" could be too low (a weight of 1 has been used, compared to a weight of 8 for the corresponding homozygous observation count). Therefore HetRank analysis was repeated with this weight increased to 4, and all other weights kept the same. (The homozygous in-house observation count still has a higher weight because if a variant can cause AOS in heterozygous form we would certainly not expect it to have been previously observed in homozygous form.)

The top 20 genes, both with and without the network-based adjustment, are shown in Table 7.15. Considering the top 20 before the network-based adjustment, we see substantial overlap of eight genes with the corresponding top 20 using the original HetRank weight parameters (given in Table 7.9). Only one of the genes had a pre-adjustment rank of over 100 using the original parameters. This is encouraging because we expect a relatively small change in the input parameters to cause only a relatively small change in the rankings. *ABCA10*, which was ranked first using the original parameters, experiences a fall in pre-adjustment rank to 270 – suggesting that the parameter change is appropriately down-ranking genes whose variants would be good candidates for disease causality were it not for previous observations in the in-house exome database. We saw a similar drop in rank, from 3 to 98, for *EXO5* – which we also noted in section 7.3.7.1 contained a variant in three exomes that almost certainly does not cause AOS because of previous observations in the in-house exome database.

The top 20 genes after adjustment are also fairly consistent with the corresponding top 20 using the original parameters (given in Table 7.10): there is an overlap of 12 genes, including the same top-ranked gene, *POLN*. No gene in the top 20 had been ranked below 51 using the original parameters. Again, this suggests a reasonable degree of robustness in the results to small changes in the input parameters. *ABCA10*'s post-adjustment rank is now 377,

Table 7.15 – Top 20 genes ranked by HetRank, with alternative input parameters, before and after network adjustment using PINA, based on 13 unsolved AOS exomes
"Prev. weights rank" = corresponding rank assigned to gene when HetRank analysis was performed using the original parameters (top 20 ranked genes without network adjustment are in Table 7.9; top 20 genes after network adjustment are in Table 7.10); "Pre-net." = rank assigned to gene without network adjustment.

Without network adjustment		
Rank	**Gene**	**Prev. weights rank**
1	*NINL*	4
2	*ZDHHC13*	2
3	*RGPD3*	10
4	*CNTNAP3B*	19
5	*PRB1*	39
6	*ANKRD36*	31
7	*CCDC144NL*	11
8	*AQR*	27
9	*AGAP6*	12
10	*KRTAP10-3*	23
11	*GJB2*	9
12	*VCX*	8
13	*ZFHX3*	83
14	*SIPA1L3*	132
15	*ABCC1*	24
16	*ZNF782*	30
17	*ATN1*	33
18	*BMP2K*	76
19	*CCDC40*	70
20	*AHCTF1*	35

After network adjustment			
Rank	**Gene**	**Pre-net.**	**Prev. weights rank**
1	*POLN*	26	1
2	*SH2D1B*	824	4
3	*KIF17*	191	3
4	*HGFAC*	1,061	6
5	*MUC4*	1,112	19
6	*LY9*	129	21
7	*DBH*	1,052	8
8	*LPA*	50	7
9	*ZNF516*	869	48
10	*SOX7*	43	20
11	*CINP*	1,833	13
12	*SCN4A*	1,310	47
13	*ADCY2*	6,868	15
14	*ZNRF4*	114	29
15	*RNF152*	4,101	36
16	*CAV3*	2,849	23
17	*FCGR3A*	501	2
18	*CD244*	8,265	24
19	*AGAP1*	190	51
20	*MST1*	31	17

having been ranked fifth using the original parameters, suggesting that the parameter change is effective.

Again we can consider common functional annotation or RGA results to establish whether the new results suggest an underlying disease mechanism for AOS.

A GO enrichment test on the top 20 genes finds no significant biological process annotation (data not shown). However, using the top 250 genes we find: "regulation of membrane potential" (16 genes; *adjP* = 0.0004) and its child term "regulation of excitatory postsynaptic membrane potential" (seven genes; *adjP* = 0.0004); "nervous system development" (51 genes; *adjP* = 0.0004), and "cell adhesion" (34 genes; *adjP* = 0.0005). Encouragingly, these are highly consistent with the enriched function that was found in the top 250 genes using the original parameters (section 7.3.7.1), as well as overlapping with the

enriched function identified in CPDBconf95 and Multinet results (section 7.3.8) and PINA results when common variants were removed (section 7.3.9.1).

When RGA is performed using alpha in the range $1 \leq \alpha \leq 250$, the most significant region is found at $\alpha = 18$ and comprises the three genes *CD244*, *LY9* and *SH2D1B* ($p < 0.0001$; see Figure 7.24). This same region was previously also found by RGA based on the HetRank results with the original parameters (see Figure 7.18b, $\alpha = 24$). The next-most significant region occurs at $\alpha = 205$ and comprises 15 genes ($p = 0.0002$; see Figure 7.24). Interestingly, this region contains the four genes that were found in the significant region at $\alpha = 61$ using the original parameters (Figure 7.18b) but also has substantial overlap with the optimal subnetworks identified by BioGranat-IG in PINA_d50 (*KIF17*, *LPA*, *LRP2*, *MAPK8IP1* and *MAP3K11* form a key part of the region in Figure 7.7, for example). While these results therefore appear promising, they do not point towards a known shared function that could explain AOS. An enrichment test on these 15 genes finds no significant GO biological process annotation. The most significantly enriched terms – including "JNK cascade" (*MAPK8IP1* and *MAP3K11*), "negative regulation of response to stimulus" (four genes), "anatomical structure morphogenesis" (six genes) and "nervous system development" (six genes) – all have a multiple-test-adjusted p-value of 0.2387. However, given that these terms relate to the broadly AOS-relevant signalling and developmental processes, and the fact that these results to some extent support what we found using BioGranat-IG, these genes may be worth studying further.

7.4 Conclusions

7.4.1 Findings from Network Analysis Regarding the AOS Disease Mechanism

AOS is a rare genetic disorder for which substantial locus heterogeneity makes it difficult to identify causal genes. We have used several network-based methods to study AOS whole exome sequencing data in light of this heterogeneity, including both candidate-gene and hypothesis-free (exome-wide) approaches.

Our results highlight sequence variants in several genes and subnetworks that could make good candidates for further study. For example:

- In section 7.3.1 the simple neighbourhood search around the known AOS gene *ARHGAP31* in the Multinet and PINAmin2 networks identified several genes of potentially relevant function (such as *ITSN1*, *KALRN*, *MAP3K11*, *SOS1* and *TIAM1*) carrying rare functional variants.

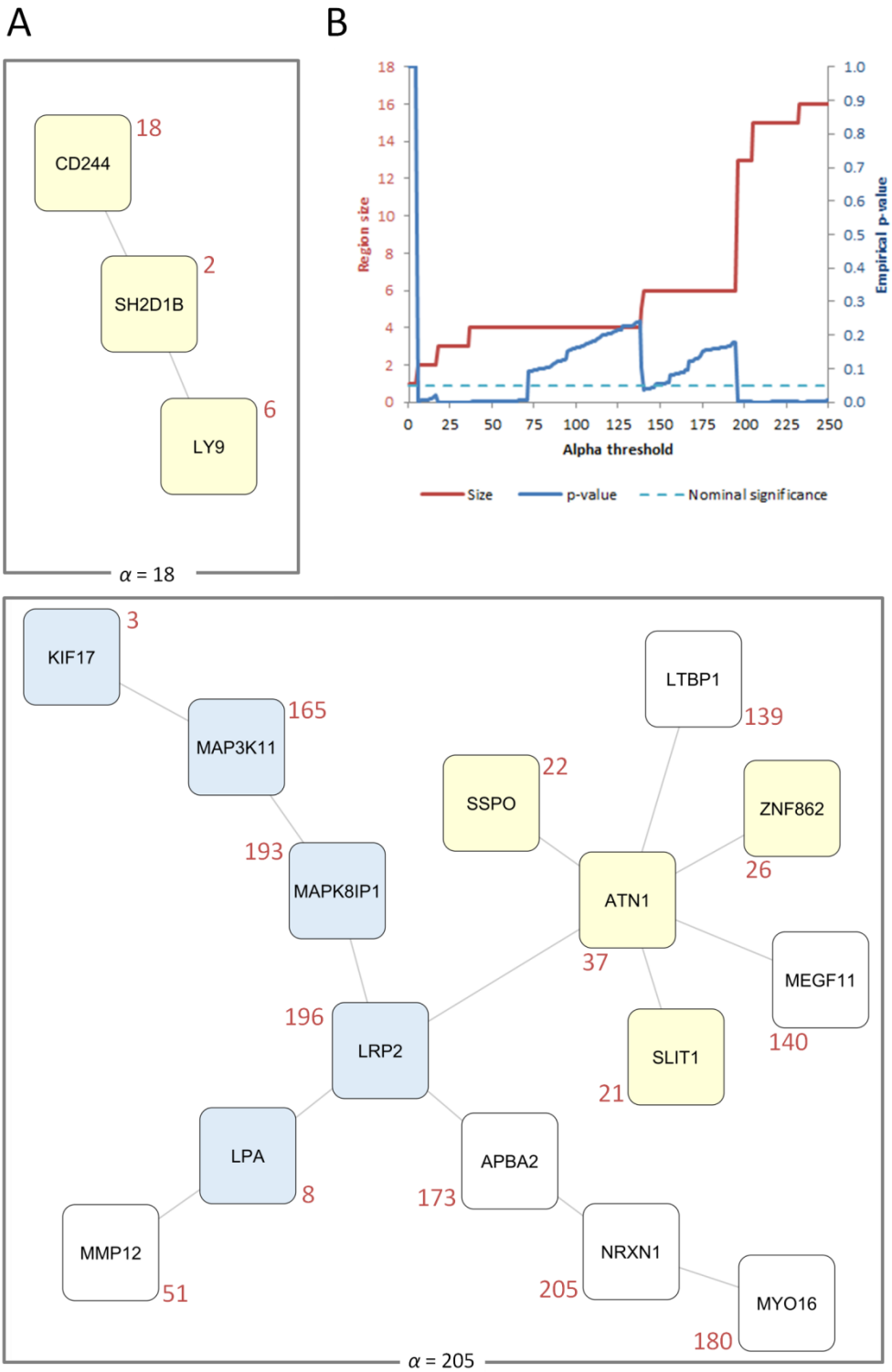


Figure 7.24 – RGA output based on PINA HetRank rankings, using alternative input parameters, for 13 unsolved AOS exomes

(A) RGA regions found at $\alpha = 98$ and $\alpha = 205$. Blue nodes indicate genes which were previously identified by BioGranat-IG in PINA_d50 (see Figure 7.7); yellow nodes indicate genes which were previously found in the RGA regions based on HetRank results using the original parameters (see Figure 7.18b). Final ranks are indicated in red. (B) Plot of region size (red) and empirical p-value (blue) by alpha value (same format as plot shown in Figure 7.18a) shows that region size is nominally significant for several ranges of alpha thresholds.

- BioGranat-IG analysis in PINA_d50 identified a small network region which comprised genes containing post-filtering variants for all of the unsolved AOS cases, was annotated by KGGSeq as significantly likely to contain a disease-causing variant, and was implicated in biological functions of potential relevance (see Figure 7.7). Genes in this region, notably *LRP2* and *MAP3K11* were subsequently found by BioGranat-IG in other networks including the high-confidence PINAmin2_d50 (see section 7.3.5.1).

- HetRank results were more difficult to interpret but results from the PINAmin2-adjusted rankings (in section 7.3.8) were of potential functional relevance, and in particular results from the PINA-adjusted rankings using the revised weighting parameters (in section 7.3.9.2) supported some of the results found by the other analyses.

It should be noted that while the existence of relevant functional annotation (such as enrichment for GO biological process terms) has been used to evaluate the results of the analyses presented here, our knowledge of gene and protein function across the genome is incomplete and constantly evolving. Therefore some of the other genes and network regions highlighted by our analyses could prove to harbour AOS-causing variants, perhaps leading to new functional insights.

Selection of genes and variants to be investigated further will be guided by geneticists with expertise in the study of AOS. The initial task will be to confirm the validity of the variants identified in these AOS-affected individuals by Sanger sequencing. Subsequently genes of interest can be examined in additional AOS cases that do not contain variants in known AOS genes, by Sanger sequencing or targeted resequencing depending on the number of samples and genes considered. A cohort of ~45 independent AOS-affected families and >50 sporadic cases is available at KCL for this purpose. Finally, genes for which multiple variants are found in the screening cohort could be selected for functional experiments. This could include study of expression levels in mutagenised versus wild-type cell lines, or gene-specific experiments in cell lines grown from AOS patient biopsies.

Of course if it were felt that the assumptions used in these analyses did not accurately reflect the current understanding of AOS genetics, the analyses could be repeated using different parameters.

It is important to recognise several limitations of the analyses performed. As alluded to earlier in the discussion of BioGranat-IG results (section 7.3.4), the incomplete nature of existing interaction networks could limit the discovery of new functional pathways underlying AOS. In addition, even for genes that are represented in the networks used,

limited whole exome sequencing coverage for some of the exomes (see Table 7.1) could result in insufficient power to detect AOS-causing sequence variants. This is especially relevant for the pair of related AOS cases (S0039 and S0301) in which only shared sequence variants were considered. In order to reduce the risk of missing a causal variant due to incomplete sequencing coverage in one of these individuals, an alternative approach could be to perform the analyses using the union (rather than the intersection) of variants identified in these cases (albeit with the disadvantage that this would introduce many confounding variants of no relevance to AOS).

Filtering levels 2-5 require that variants' zygosities match the expected mode of inheritance for each AOS case. Incorrect judgements about the mode of inheritance for each AOS case could therefore render these assumptions inappropriate, and a further source of error could be incorrect zygosity-calling for sequenced variants (zygosity being assigned by likelihood based on the observed reads). Note also that although no filtering was undertaken for the HetRank analyses presented here, zygosity was used as a ranking criterion.

AOS is a rare disease with severe phenotype, and as such is expected to be caused by a rare highly-penetrant variant. As with any whole exome sequencing study for such a disease, the number of variants per exome remaining after filtering could be reduced by increasing the number of unaffected exomes used for comparison. This would be of particular benefit for the simple neighbourhood search and BioGranat-IG analyses. Therefore, it may be informative to repeat the analyses at a future time when a larger in-house exome database is available for filtering.

Finally, AOS is known to have substantial clinical variability (Snape et al. 2009), and phenotypic differences were observed in the cases studied here (see Table 7.1). While the analyses presented here are consistent with a hypothesis that genetic heterogeneity could to some extent explain this phenotypic heterogeneity (indeed, they test previous assertions that interaction networks could be used to elucidate this relationship (Oti and Brunner 2007)), we have not made use of any specific phenotype information. Future analyses might therefore make use of tools that perform phenotype-aware prioritisation of whole exome sequencing variants, such as PHIVE (PHenotypic Interpretation of Variants in Exomes; Robinson et al. 2013).

7.4.2 Relative Merits of the Network-Based Methods

We have employed several network-based methods to address the problem of genetic heterogeneity in a rare monogenic disease. Due to its rarity and established locus heterogeneity, AOS is a good match to the genetic model assumed by BioGranat-IG and

HetRank; the fact that there are several known causal genes as well as many unsolved cases makes it an ideal disease for the application and comparison of these methods.

The existence of known causal AOS genes makes the simple neighbourhood search around these genes a natural starting point. It is effectively a candidate-gene approach; for any promising variants found by this search it should (in theory) be relatively easy to establish a functional link to an existing AOS disease mechanism due to network proximity. For the same reason it is encouraging when the subsequent hypothesis-free approaches identify some of the same genes (e.g. as we saw with *MAP3K11* here; see section 7.3.4.1).

In general, however, overlap between the results obtained using different analysis methods should not be unquestioningly taken as confirmation that we have an interesting finding. This is because all of the methods will be predisposed to highlight genes containing rare non-synonymous variants in multiple AOS cases, such as those listed in Table 7.5. In addition, a degree of overlap between the results obtained using different networks might also be anticipated, since there is some degree of overlap between the networks themselves (indeed, four of the five networks used are partially or entirely based on PPIs from common underlying databases) (see chapter 2 and Table 2.3 in particular). In one sense, then, it is reassuring that genes such as *LRP2* and *NINL* occur in numerous optimal and near-optimal BioGranat-IG subnetworks and are highly ranked by HetRank across multiple networks; however, these methods can only prioritise genes for further study and a true role in the AOS disease mechanism can only be established by additional investigations that are independent of the interaction networks.

As discussed in section 7.3.5.1, limited network coverage coupled with the abundance of sequence variants identified by whole exome sequencing tends to cause BioGranat-IG's heuristic (minimum distance and multi-minimum distance) searches to report optimal subnetworks containing many false positive findings. This is due to these searches' mandate to continue growing subnetworks until variants have been found for all exomes (or as many exomes as possible). Perhaps the most useful results from BioGranat-IG are therefore generated by the exact triplet and quadruplet searches, which can identify relatively small subnetworks with a concentration of post-filtering variants in different exomes.

It is disappointing that in none of the networks were the true causal genes found in optimal BioGranat-IG triplets and quadruplets at filtering levels 1 and 2 (where variants in the six solved AOS exomes were filtered no differently from the unsolved cases), and particularly so at level 3 (where the input gene lists for the six solved cases were fully

filtered to include only the causal genes).* In networks where the known AOS genes are present, we do not find them in the optimal triplets and quadruplets because more concentrated enrichment of post-filtering variants occurs in other parts of the networks. An alternative way to look at this is that because the genetic heterogeneity is so great (at least in the AOS cases studied here), and the known AOS genes are insufficiently proximal in the networks, BioGranat-IG is underpowered to recover these genes.

BioGranat-IG analyses were performed using hub-free versions of the networks because BioGranat-IG results tend to be biased by the presence of hub genes. Hub-removal resulted in the loss of known AOS genes from several networks (as listed in Table 7.7). This meant firstly that BioGranat-IG was unable to recover these genes for the solved AOS cases at filtering levels 1-3, but secondly that the searches at filtering level 3 in particular were unable to be implicitly weighted towards these genes by the variants in solved cases, in order to propose potential extensions of known AOS disease pathways.

Further, network hub removal as used here removes highly-connected genes using an arbitrary degree threshold. This can have a considerable impact on the optimal subnetworks found by BioGranat-IG. For example, *LRP2* was identified as a key gene in several subnetworks in PINA_d50 (see for example Figure 7.7) but has a degree of 49 in the full PINA network. If the threshold for hub removal had been any lower than 50 as used here, this gene would have been removed. A more sophisticated strategy for hub removal is suggested in the Concluding Discussion (section 9.2.3).

HetRank analysis achieved mixed success. It was able to rank *NOTCH1* seventh when solved AOS cases were included (using PINA for the network-based rank adjustment; see Table 7.12); this largely resulted from true AOS-causing variants, in the gene itself for two of the solved cases and in its direct neighbour *RBPJ* for a third. We also found some promising results using the unsolved AOS cases only, as described in the previous section. In section 7.3.9.2 we saw that by careful consideration of initial HetRank results, the weighting parameters could be adjusted to more accurately reflect expectations about the properties of AOS-causing variants, resulting in a more plausible final ranking.

On the other hand, however, while HetRank was able to prioritise *NOTCH1* it did not find its neighbour *RBPJ*, confirming (as we found using simulated exome data in chapter 5) that its performance is limited at high levels of genetic heterogeneity. HetRank results were often difficult to interpret and a comparison of results across multiple networks

* Note however that at filtering level 3: *NOTCH1* was found in a near-optimal triplet in PINAmin2_d50; *NOTCH1* and *RBPJ* were found in near-optimal quadruplets in PINAmin2_d50 and CPDBconf95_d50; and *DOCK6* and *RBPJ* were found in near-optimal quadruplets in COXPRES30_d50.

suggests that the method relies too heavily on indirect evidence (via interaction data) and insufficiently on direct evidence (via the variants contained in the genes themselves) for highly-ranked genes (see section 7.3.8). HetRank's design takes into account node connectivity when adjusting gene ranks, but as illustrated by Figure 7.17 the approach needs to be refined in order to adequately deal with the problem caused by hub genes. Likewise as discussed in section 7.3.7.1 subsequent iterations of the tool should address the fact that genes not represented in the network are at a considerable disadvantage in the final rankings. The fact that the results in section 7.3.9.1 differ markedly from those found earlier in the chapter (see Table 7.14) also suggest that the many variants unlikely to be of relevance to the disease in question (in the case of AOS, common variants) are too influential over the final rankings. For ranking, rather than filtering, of variants to be a viable analysis method this problem will also need to be addressed.

Given some of these difficulties with the application of HetRank, in particular the somewhat volatile effect of the network-based rank adjustment, there are immediate alternative approaches that could be considered to make use of the HetRank framework in a more tractable manner. For example, HetRank could be used without a network-based adjustment to produce a final gene ranking that reflects only the direct evidence for disease-causality across all exomes in the study; RGA could then be used to address locus heterogeneity by identifying small network regions enriched for highly-ranked genes. Alternatively, HetRank could be used only for the purpose of ranking genes within individual exomes; BioGranat-IG analysis could then be performed using the genes exceeding some fixed threshold as the candidate genes for each exome.

The network-based analyses presented in this chapter have proposed several novel disease pathways that can be investigated further. However, it is clear that none of these methods can fully replace the knowledge and experience of a skilled geneticist who has studied the disease closely. The tools are thus limited to prioritising variants and genes for a genetic researcher, and cannot directly demonstrate disease involvement. However, given the wealth of data produced by NGS methods this function can still be of great value.

8 Analysis of Familial Crohn's Disease Exome Sequence Data using Network Methods

8.1 Introduction

8.1.1 Background

Crohn's disease (CD) is one of the two main types of inflammatory bowel disease (IBD), the other being ulcerative colitis (UC). IBD is a group of disorders characterised by relapsing chronic inflammation of the gastrointestinal tract and which can result in diarrhoea, bleeding, weight loss, abdominal pain and fatigue (Mathew 2008; Matricon et al. 2010). CD is distinguished from UC by the location and nature of inflammation. While CD inflammation usually occurs in the ileum and colon, it can affect any part of the gastrointestinal tract (unlike UC, where it is confined to the colon and rectum); it is also transmural (whereas UC is limited to the intestinal lining) (Mathew 2008; Matricon et al. 2010). CD is a relatively common disorder; prevalence estimates range from 26 to 375 cases per 100,000 people in populations of European ancestry (Loftus 2004; Mathew 2008).

CD is a complex disease for which genetic, environmental and lifestyle factors all contribute to disease risk. The microbial environment within the gut is thought to play a major role in CD pathogenesis, while smoking is an example of a lifestyle factor that can increase disease susceptibility and severity (Matricon et al. 2010; Khor et al. 2011). However, among complex diseases CD appears to have a relatively strong genetic component (Mathew 2008; Khor et al. 2011; Liu and Anderson 2014; Zhang and Li 2014). This is implied by an estimated concordance rate of around 30.3% in monozygotic twins (Brant 2011) and a sibling relative risk ratio in the range 15-42 (Halme et al. 2006).

Much progress has been made in understanding the genetic basis of CD, with the first susceptibility gene, *NOD2*, being identified by linkage-informed candidate-gene approaches in 2001 (Hugot et al. 2001; Ogura et al. 2001). Since then, the advent of genome-wide association studies (GWAS) has heralded a rapid increase in the number of genomic loci associated with CD risk, with a 2010 meta-analysis bringing this number to 71 (Franke et al. 2010). In 2012, Jostins *et al*. reported a GWAS meta-analysis that combined CD and UC cases and brought the total number of loci implicated in IBD to 163. A combined IBD association test was performed, along with separate tests for CD and UC; in total 30 CD-specific associated loci were found, and 76 IBD-associated loci also achieved

genome-wide significance ($p < 5\times10^{-8}$) in the CD test (Jostins et al. 2012). The associated loci are estimated to explain 13.6% of disease variance in CD, with the biggest contribution coming from variants in *NOD2*, followed by *IL23R* (Jostins et al. 2012).

Despite the number of associated loci, the molecular mechanisms underlying CD are still not fully understood, although several functional pathways are thought to play a role in pathogenesis. Firstly, *NOD2* encodes a protein which plays an important role in the innate immune system by recognising bacterial molecules and activating NF-κB, a rapid-acting primary transcription factor that stimulates the immune response. One model for CD suggests that loss-of-function mutations in *NOD2* can impair this immune response leading to increased bacterial survival (Mathew 2008). (Although some CD-associated variants in *NOD2* appear not to affect NF-κB activation, implying that impairment of other functions of the NOD2 protein can also influence CD susceptibility (Rivas et al. 2011)). Secondly, the IL23 pathway is thought to be a key process influencing CD through its role in mucosal inflammation. Variants in the receptor gene *IL23R* have been found to be protective against CD, while others appear to elevate risk (Duerr et al. 2006; Rivas et al. 2011). The IL23 cytokine activates proinflammatory Th17 cells, and several proteins involved in the downstream signalling pathway are also encoded by genes associated with CD (Brand 2009). Thirdly, associated variants in autophagy-related genes (*ATG16L1* (Hampe et al. 2007) and *IRGM* (Parkes et al. 2007)) suggest a role for this process, possibly by impairing its ability to remove intracellular microbes (and note also that autophagy can be activated by NOD2, raising the potential of interactions with another CD pathway) (Khor et al. 2011).

Many of the associated loci reported in the recent GWAS meta-analysis supported these disease models, while others suggested that there are additional mechanisms of CD biology yet to be elucidated (Jostins et al. 2012). Interestingly, several network-based methods were employed in this study to explore potential IBD pathways: GRAIL and DAPPLE were used to prioritise genes around associated loci (using text-mined interactions and protein-protein interactions respectively; see introduction section 1.4.6), while weighted gene co-expression network analysis was used to highlight associated modules (see section 1.4.4). This is encouraging because it suggests that in the case of IBD (and CD in particular) interaction networks can help to bridge the gap between hypothesis-free genetic observations and a full understanding of the molecular processes that result in the clinical phenotype.

As a complex disease, most cases of CD are expected to be caused by a combination of genetic, environmental and lifestyle factors, and most genetic risk factors have small effect size (Jostins et al. 2012). This multifactorial nature of CD means that despite the relatively increased disease risk carried by relatives of CD patients, strong patterns of recurrence within families are atypical. For families where such patterns occur it is therefore

reasonable to consider alternative models for the genetic basis of CD; one possibility is that for some such families one or a small number of inherited rare and highly-penetrant variants could be responsible for most of the disease risk, making these cases effectively monogenic or oligogenic forms of CD. There is precedent for this model of disease because several other complex traits, such as diabetes and hypertension, are known to have rare monogenic forms (Peltonen et al. 2006). Interestingly, there are also several monogenic forms of very early-onset IBD (Uhlig et al. 2014). Although these disorders generally present in infancy and the clinical features are not identical to adult-onset CD, it nonetheless seems feasible that monogenic forms of adult-onset CD could exist and give rise to familial recurrence.

Using whole exome sequence data from 25 pedigrees exhibiting recurrent CD, this chapter will test the hypothesis that families with multiple affected members are more likely to carry one, or very few, highly-penetrant risk variants. As will be seen, no single gene is clearly implicated in these familial cases of CD and so network methods will be employed to examine the possibility that familial CD has a monogenic or oligogenic basis with genetic heterogeneity.

8.1.2 Analysis Strategy

Twenty-four pedigrees in which three or more individuals have CD, and one pedigree in which two individuals have CD, were recruited for study at King's College London (KCL) or at the Institute of Clinical Molecular Biology (ICMB) at the University of Kiel, Germany. For each pedigree between two and four affected individuals were whole exome sequenced (67 individuals in total). Under the assumption that instances of CD within each family are caused by the same variant(s), we sought candidate disease-causing variants shared by all sequenced exomes in a pedigree. Subsequently, these shared variants will be referred to as variants *carried by a pedigree*.

Three independent network methods were applied to the exome data to propose familial CD genes. All three methods are hypothesis-free in the sense that they use no prior knowledge of CD aetiology. Firstly the BioGranat-IG tool developed in chapter 4 was used to search for subnetworks in which all or most pedigrees carry a shared post-filtering variant, testing the possibility that these CD cases are monogenic but subject to locus heterogeneity (see Figure 8.1a).

To test whether an oligogenic basis for CD could be established in any of the pedigrees by identifying multiple post-filtering variants in functionally-related genes, a constrained version of Region Growing Analysis (RGA; described fully in chapter 6, section 6.2) was employed separately for each pedigree (see Figure 8.1b). This method will be referred to as *within-pedigree RGA*. The relevance of any regions found was examined by

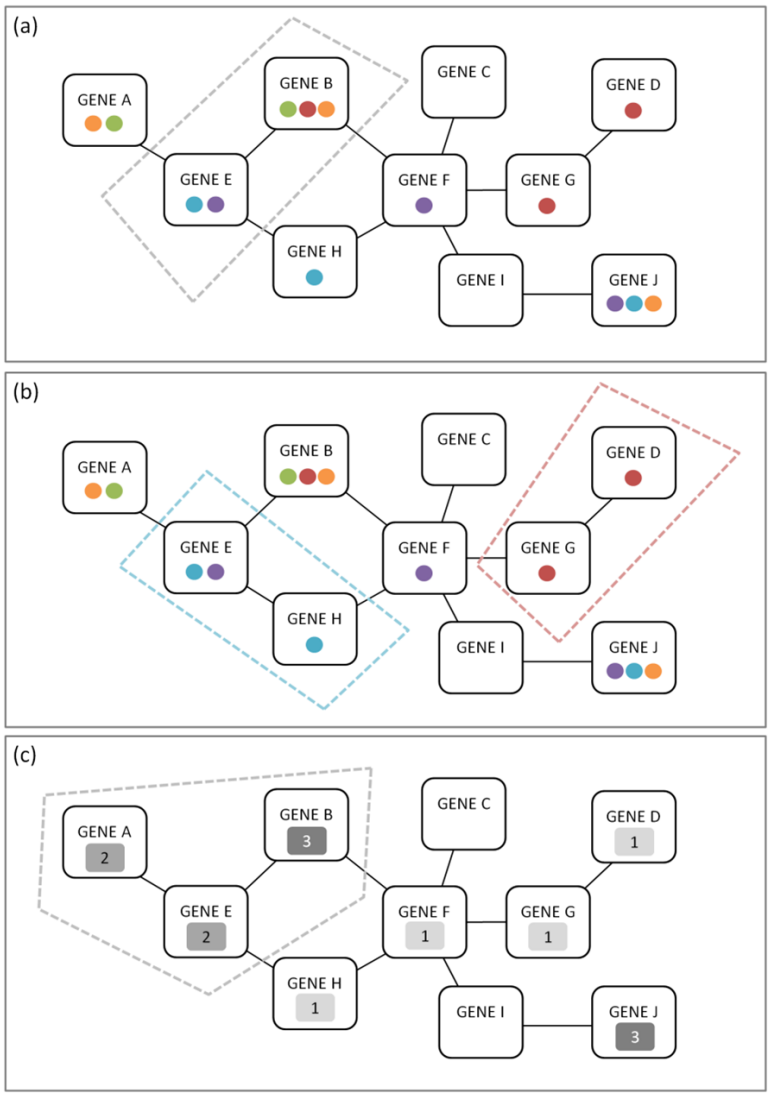


Figure 8.1 – Network-based methods used to analyse CD exome data
(a) BioGranat-IG: different-coloured tokens signify post-filtering variants in different pedigrees. The tool seeks a small connected set of genes in which all or most pedigrees carry a variant. (b) Within-pedigree RGA: connected sets of genes containing post-filtering variants in the same pedigree (tokens of the same colour) are sought. This method hypothesises that the variants in such genes could underlie an oligogenic form of familial CD. (c) Across-pedigree RGA: no assumptions are made about the number of variants expected to be observed in each pedigree; RGA is performed on a list of genes ranked by the number of pedigrees in which they carry a post-filtering variant.

testing for statistically significant clustering of regions within the network (making the broad assumption that there will be an overlap between a topological module in the network and a so-called CD disease module (Barabasi et al. 2011)).

For the third method, no assumptions are made about the number of causal variants that we expect to identify for each pedigree. Instead we use RGA to test for network regions that are enriched for post-filtering variants in any pedigree (see Figure 8.1c). This will be referred to as *across-pedigree RGA* and is a more typical use of the RGA tool, whereby any

significantly large enriched network regions can be examined as candidate disease pathways (Lehne 2011).

8.2 Methods

8.2.1 Exome Data

Whole exome sequencing was performed on 22 individuals with CD from 9 pedigrees recruited at KCL, and 45 individuals with CD from 16 pedigrees recruited at ICMB, according to the procedure described in chapter 2, section 2.3. Pedigree structures, affection statuses and sequenced individuals are illustrated in Figure 8.2.

In two of the pedigrees, one individual was affected with UC. In a small number of the pedigrees individuals affected or unaffected with CD displayed other inflammatory or immune-mediated phenotypes, such as asthma and psoriasis (data not shown). This additional phenotype data was not considered further in this study.

Table 8.1 summarises the sequenced exomes for the 25 pedigrees. At 20× coverage, exome capture ranged from 63.8% to 92.6%, with only three of the 67 exomes falling below 80%. This is generally considered to be a reasonable level of coverage for variant calling, although 90% or higher is desirable to maximise the power to accurately detect sequence variants (Mertes et al. 2011).

338 unaffected control exomes (a subset of the exomes covered by the in-house exome database which will be referred to subsequently as *non-CD control exomes*) were also used in the analysis. These represent unrelated individuals of European ancestry sequenced as part of the KCL rare disease programme. As such they may include causal variants for a range of diseases other than CD. Since these diseases are clinically diverse the data are considered suitable for use as controls for variant filtering. These exomes will also contain known or as-yet-unidentified risk variants having small effects on susceptibility for the common complex form of CD. This is acceptable because we are specifically interested in high-penetrance variants with large effect in this study.

8.2.2 Variant Filtering

Since we are testing a mono/oligogenic disease model, all three analyses require a set of candidate causal variants for each pedigree. Firstly, it was assumed that within each pedigree one or a small number of identical-by-descent variants cause CD in all affected individuals. Therefore for each pedigree the set of variants shared by all sequenced CD cases was obtained. The number of variants for each pedigree is given in Table 8.1. Due to the

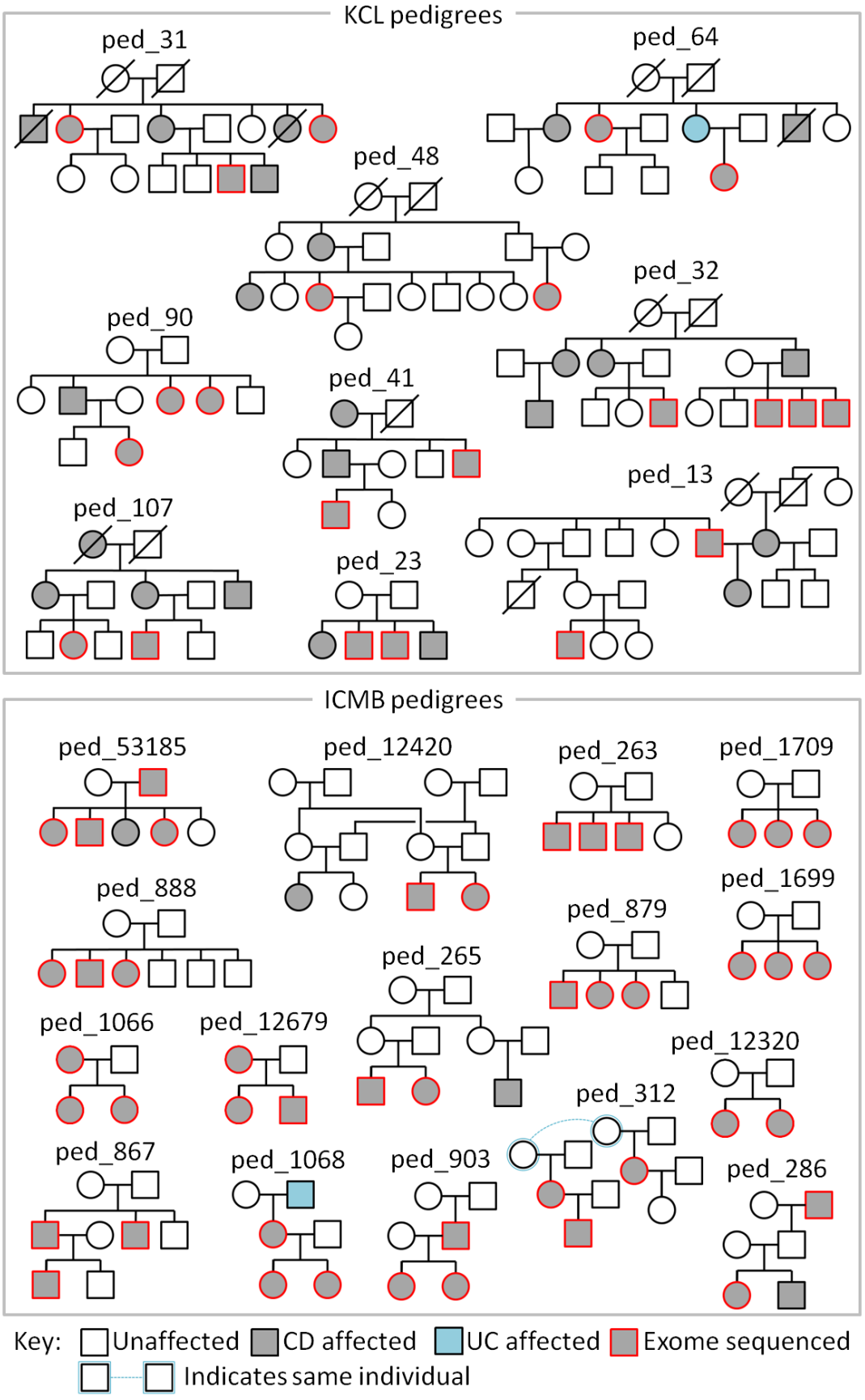


Figure 8.2 – Familial CD pedigree diagrams
Affection status for CD and UC are indicated.

Table 8.1 – Summary of sequencing for CD pedigrees
"Min-" and "Max coverage (to 20×)" refer to minimum and maximum coverage across all sequenced exomes within a pedigree; "relaxed filtering" refers to variant-filtering performed for within-pedigree RGA analysis.

Pedigree ID	# Exomes sequenced	Min coverage (to 20×)	Max coverage (to 20×)	# shared variants	# variants after filtering (in # genes)	# variants after relaxed filtering (in # genes)
KCL pedigrees:						
13	2	90.8%	91.2%	11,122	53 (53)	191 (175)
23	2	81.0%	81.2%	14,983	115 (107)	410 (373)
31	3	80.2%	83.0%	8,932	34 (34)	99 (95)
32	4	63.8%	83.9%	5,729	7 (7)	23 (22)
41	2	80.4%	83.9%	11,173	47 (47)	152 (142)
48	2	81.5%	83.0%	10,715	35 (34)	126 (116)
64	2	80.4%	81.6%	11,833	54 (54)	205 (188)
90	3	81.9%	83.0%	8,754	36 (36)	118 (110)
107	2	81.2%	82.3%	10,748	30 (30)	118 (108)
ICMB pedigrees:						
263	3	79.5%	92.0%	10,921	53 (53)	197 (175)
265	2	90.3%	91.3%	15,208	138 (131)	360 (326)
286	2	87.9%	91.2%	11,950	49 (49)	168 (156)
312	3	90.2%	92.3%	8,767	36 (35)	123 (115)
867	3	82.3%	91.7%	9,899	60 (60)	165 (160)
879	3	92.5%	92.6%	12,836	102 (101)	290 (265)
888	3	90.7%	91.9%	12,701	106 (98)	309 (272)
903	3	89.9%	91.6%	10,498	70 (68)	201 (176)
1066	3	85.5%	92.0%	15,439	139 (129)	428 (382)
1068	3	66.6%	90.6%	10,260	59 (59)	214 (201)
1699	3	86.9%	92.5%	12,576	84 (84)	259 (227)
1709	3	89.4%	92.5%	12,380	105 (104)	273 (248)
12320	2	89.6%	90.1%	16,223	173 (170)	470 (428)
12420	2	90.5%	91.9%	15,520	126 (120)	406 (366)
12679	3	82.5%	86.3%	10,473	64 (64)	200 (182)
53185	4	84.3%	92.6%	8,256	52 (50)	120 (112)

varying number and relatedness of sequenced cases between pedigrees, there is a wide variation in the number of shared variants per pedigree, from 5,729 to 16,223.

Subsequently, variants carried by each pedigree were filtered to exclude from consideration those less likely to cause a rare monogenic disease. Synonymous variants were excluded. No assumptions were made about the expected zygosity for causal variants. Since we expect causal variants to be rare and highly-penetrant, heterozygous variants were required to have alternative allele frequencies of <1% according to both 1000 Genomes

Project and Exome Variant Server (EVS) annotation, and to have been observed fewer than ten times in heterozygous form (and never in homozygous form) in the in-house exome database (approximately 850 exomes in total). Homozygous variants were required to have alternative allele frequencies of <10% (corresponding to a genotype frequency <1%), and fewer than ten previous observations in homozygous form in the in-house exome database. Finally, variants were filtered at the gene level against the 338 non-CD control exomes in order to address the problem that large and highly-polymorphic genes are more likely to be falsely implicated by variant-filtering approaches (as discussed in section 1.3.5). Variants were excluded in any gene in which 20 or more non-CD control exomes, after being subject to the same filtering steps, carried a post-filtering variant. The number of variants per pedigree after filtering is given in Table 8.1, and ranges from 7 (all in different genes) to 173 (in 170 unique genes).

For within-pedigree RGA only, where subnetworks are identified in which the same pedigree carries two or more variants, filtering was relaxed to allow for the decreased probability that a member of a population carries two variants together. In this case, heterozygous variants were required to have alternative allele frequencies of $<\sqrt{1\%}$ (i.e. <10%) according to both 1000 Genomes Project and EVS annotation, and to have been observed fewer than 100 times in heterozygous form (and never in homozygous form) in the in-house exome database. Homozygous variants were required to have alternative allele frequencies of $<\sqrt{10\%}$, and fewer than 100 previous observations in homozygous form in the in-house exome database. Variants were excluded in any gene in which 80 or more of the non-CD control exomes, after being subject to the same filtering steps, carried a post-filtering variant (since $80/338 \approx \sqrt{20/338}$). Again, Table 8.1 gives the number of variants per pedigree after filtering, which ranges from 23 (in 22 unique genes) to 470 (in 428 unique genes).

We make the assumption to ignore any potential population structure differences between the British (KCL) and German (ICMB) pedigrees. This is because in general we are seeking to prioritise single variants that could cause CD, rather than to perform statistical association testing; a rare non-synonymous variant in a gene is assumed to be equally likely to have a pathogenic effect in both populations. It is possible therefore that since non-CD control exomes are predominantly drawn from a British population, gene-level filtering against controls could allow through variants in ICMB pedigrees in a small number of genes that would potentially have been excluded had the non-CD control exomes been drawn from a German population, leading to a few false positive findings in ICMB pedigrees. As will be seen in the results section 8.3.1, the use of a different sequencing platform for ICMB pedigrees relative to non-CD control exomes appears more likely to cause this problem.

Finally note that known *NOD2* risk variants were identified in several of the pedigrees, including two in pedigrees 13 and 12420 that pass the (non-relaxed) frequency-filtering criteria described above (but were subsequently excluded because *NOD2* is relatively highly polymorphic, containing post-filtering variants in 35 of the 338 non-CD control exomes). Since *NOD2* is well-studied in the context of CD, and these variants are known to increase risk as part of a multifactorial mechanism (rather than being highly-penetrant disease-causing variants), these pedigrees were not excluded from the network analysis. It is hypothesised that in these strongly familial cases there may be other rare highly-penetrant variants that are chiefly responsible for CD.

Likewise, none of the pedigrees were screened for variants in genes known to cause very-early-onset IBD (Uhlig et al. 2014). This is because a different disease mechanism is expected to be responsible for adult-onset CD in these pedigrees. However, where relevant any very-early-onset IBD genes that are identified by our analyses will be noted.

8.2.3 Interaction Networks

Analyses were performed using several of the networks described in detail in chapter 2. These comprised three protein interaction networks (PINs; PINA_d50 and the two smaller but higher-confidence networks PINAmin2_d50 and CPDBconf95_d50), a co-expression network (COXPRES30_d50) and a network integrating several types of interaction data (Multinet_d50). These are the hub-free versions of the networks because hub genes can be overrepresented in optimal BioGranat-IG subnetworks, and can reduce the power of RGA (Lehne 2011).

To ensure consistency between network node labels and exome gene symbols, a list of gene symbol synonyms was obtained from the HUGO Gene Nomenclature Committee (Gray et al. 2013). An R programme was implemented to replace the node label with a synonymous symbol for any network node whose label did not map to any gene in the exome data, but for which an unambiguous mapping existed for one of the synonyms. However, in practice no changes were necessary for any of the networks.

8.2.4 BioGranat-IG Analysis

BioGranat-IG analysis, as detailed in chapter 4, was performed for all five (hub-free) networks using gene lists derived from the post-filtering variants carried by each pedigree.

In chapter 7, discussion of the optimal subnetworks for Adams-Oliver syndrome (AOS) highlighted that BioGranat-IG's heuristic minimum distance and multi-minimum distance searches frequently introduce false positive findings into optimal subnetworks, due to their tendency to keep extending a subnetwork until it contains variants for as many

exomes as possible. For CD, then, only the (exact) triplet and quadruplet searches are undertaken, which examine how many of the pedigrees could potentially have CD due to three or four interacting genes. The searches were repeated for each network using "optimal" (size flexibility = 0; number flexibility = 0) and "near-optimal" (size flexibility = 1; number flexibility = 1) parameter sets. Where several overlapping optimal triplets or quadruplets are identified, they can be considered separately or merged into a region. These searches should therefore highlight network regions enriched for post-filtering variants that cover as many pedigrees as possible.

To prioritise alternative optimal subnetworks, KGGSeq-prioritisation is used, as described fully in chapter 6, section 6.1. This method tests whether the variants in a given subnetwork are more likely to include one which causes a monogenic disease, as quantified by the KGGSeq variant effect prediction tool (Li et al. 2012), than a set of variants chosen uniformly at random from the same pedigrees without regard to network connections between genes.

8.2.5 Within-Pedigree RGA

To test each pedigree for multiple variants in connected genes, which might indicate an oligogenic disease mechanism, RGA (described fully in chapter 6, section 6.2) was performed separately for each pedigree. RGA is designed to identify connected regions of a network that are enriched for highly-ranked genes from a ranked gene list provided as input. However, RGA can also be used in "binary" form by providing an input gene set (without ranks) and having RGA report connected regions of a network made up of genes in this set (or optionally regions where genes in the set are connected via "jumps" of up to one node). The advantage of this approach is that RGA's degree-constrained network permutation test can be used to estimate how likely it is that the largest identified region could have occurred by chance.

For each pedigree, the gene set used as input comprises all genes containing post-filtering variants for that pedigree (using the relaxed filtering criteria described in section 8.2.2 to allow for the decreased probability that a member of a population carries two variants together). The binary form of RGA is performed using all five (hub-free) networks. Jumps are not permitted. 10,000 degree-constrained permutations of each network are used to establish whether a significantly large region is identified for any pedigree, with RGA's default standard deviation of 5.0 nodes used for label-shuffling.

For each network tested, the regions found are compared across all pedigrees. Of particular interest is whether the largest regions found in different pedigrees (whether or not they are significantly large) cluster in one or more parts of the network. If so, this could

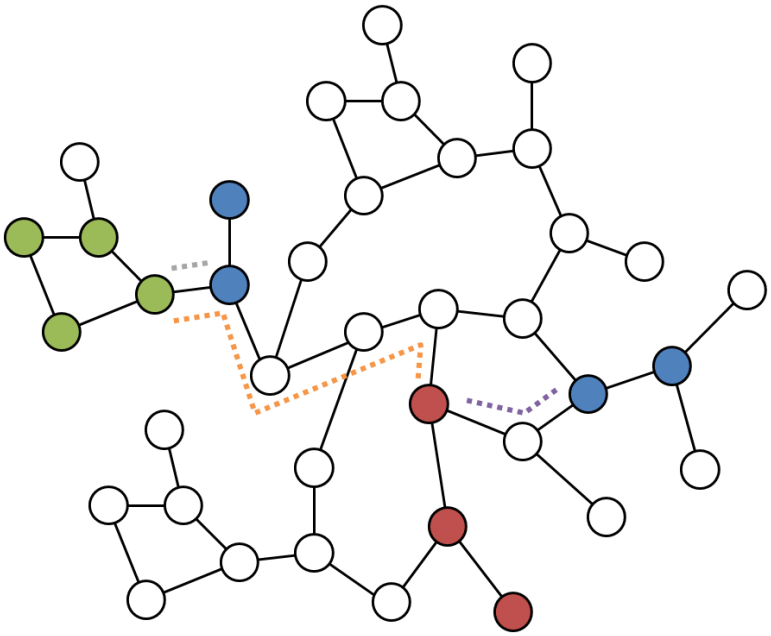


Figure 8.3 – Region clustering measures
Largest regions are illustrated for three pedigrees *G*, *R* and *B*. Pedigree *G* has a largest region of four genes ($m_G = 4$; green nodes); pedigree *R* has a largest region of two genes ($m_R = 3$; red nodes); pedigree *B* has a largest region of two genes and there are two such regions ($m_B = 2$, $l_B = 2$; blue nodes). For the first clustering measure, the distances between the regions are calculated as: *G* to *R* has distance 5 (indicated by orange dotted line); *G* to *B* has distance 1 (grey dotted line); *R* to *B* has distance 2 (purple dotted line). The average distance is (5 + 1 + 2)/3 = 2.67. For the second clustering measure, of the three pairs of pedigrees (*G*+*R*, *G*+*B* and *R*+*B*) only one has overlapping or adjacent regions (*G*+*B*; grey dotted line).

suggest that the genes in that part of the network are part of a functional module of relevance to CD.

To quantify the evidence for clustering, only regions found within the large component of each network are considered. For the network *G*, suppose that pedigree *i* has post-filtering variants in n_i genes in the large component of *G*, and that the largest region found using RGA contains m_i genes. If pedigree *i* is found to have $l_i > 1$ such regions then all $l_i \times m_i$ genes are included together for pedigree *i* for the purposes of measuring clustering. The observed degree of clustering in network *G* is measured using two methods (illustrated in Figure 8.3). Firstly, the average distance between largest regions is calculated, where the average is taken across all pairs of pedigrees for which regions were identified. Here, the distance between regions for pedigrees *i* and *j* is the length of the shortest path connecting any of the $l_i \times m_i$ genes in pedigree *i*'s largest regions in *G* to any of the $l_j \times m_j$ genes in pedigree *j*'s largest regions in *G*. The distance is zero for pedigrees with overlapping largest regions. Secondly, since this first method uses regions identified in all pedigrees and we are also interested in clustering among a subset of pedigrees, the number of pairs of pedigrees having overlapping or adjacent regions (i.e. a distance ≤ 1) is counted.

In order to test the null hypothesis that the observed degree of clustering could occur by chance due to the number and size of regions compared, the degree of clustering was also determined in 10,000 sets of randomly-generated regions (having the same number and size as the observed regions for each pedigree). To avoid potential bias due to topological properties (such as node degree or cliques) that could predispose certain parts of the network to contain regions, these regions were generated using an acceptance-rejection approach that emulates RGA as follows. For pedigree i, n_i genes were uniformly randomly selected from the large component of G. The largest regions formed by these nodes were identified. If these regions had size m_i, and $\geq l_i$ such regions were found, then these were accepted as a sampled observation (with l_i of the regions chosen at random in cases where $> l_i$ were found). Otherwise the regions were rejected. This process was repeated until 10,000 sampled observations had been accumulated for each pedigree. A p-value for the test that clustering occurs by chance was generated as the fraction of the 10,000 sampled sets of regions that exhibited an equal or greater degree of clustering (using the two measures described previously).

Randomly-generated regions were produced using a custom bundle in BioGranat (see chapter 2); subsequent analysis was performed in R (R Development Core Team 2013).

8.2.6 Across-Pedigree RGA

Finally, across-pedigree RGA was performed to identify network regions enriched for post-filtering variants in any pedigree, with no assumptions made about the underlying genetic architecture for each pedigree. This is done by ranking genes according to the number pedigrees in which they contain post-filtering variants, and using RGA to identify regions enriched for highly-ranked genes.

Ranked gene lists were obtained using two methods. Firstly, genes were ranked using a *simple count*. For each gene the number of pedigrees carrying a post-filtering variant was established (filtering is described in section 8.2.2; note that the relaxed criteria for variant filtering, as used for within-pedigree RGA, were not used here). Genes were ranked according to this count, and any gene in which 20 or more of the 338 non-CD control exomes carried a post-filtering variant was excluded from the list (meaning it is unranked and will not be a member node for any region; see chapter 6, section 6.2). For reasons that will be described in the results section 8.3.1 below, it was also necessary to exclude the four top-ranked genes.

Secondly, since the simple count ranking has limited ability to discriminate between genes (all but 16 genes contain post-filtering variants in zero, one or two pedigrees[*]), a more granular ranking was obtained using a *case-control* approach. Filtering was performed as described in section 8.2.2, except that all genes were retained (genes in which 20 or more of the 338 non-CD control exomes carried a post-filtering variant were not excluded). Genes were subsequently ranked according to the degree to which the number of pedigrees carrying a post-filtering variant exceeded the number that would be expected assuming that the distribution of post-filtering variants among genes in the 338 non-CD control exomes is typical. Thus:

- Genes in which post-filtering variants were found only in the case pedigrees and not in the non-CD controls were ranked first, ordered by decreasing number of pedigrees carrying a variant.

- Genes in which both case pedigrees and non-CD control exomes carried a post-filtering variant were then ranked according to:

$$\frac{O_g - E_g}{\sqrt{E_g}},$$

where O_g gives the number of pedigrees in which gene g carries a post-filtering variant, and:

$$E_g = N_g \times \frac{\sum_{h \in H} O_h}{\sum_{h \in H} N_h}$$

gives the expected number. Here N_g is the number of the 338 non-CD controls carrying a post-filtering variant in gene g and the sums are taken across the set of genes H which contain post-filtering variants in both the case pedigrees and non-CD control exomes (the ratio of sums ensures the correct total number of expected observations).

The same four genes were excluded for case-control ranking as for the simple count ranking. (Note that as discussed in section 8.2.2 above, we are ignoring any potential population structure differences between German [ICMB] pedigrees and British [KCL] pedigrees and non-CD control exomes. For this analysis specifically, population structure and sequencing platform differences could lead to a subtle skewing of the rankings because

[*] This drops to 12 genes after removing the four top-ranked genes.

the expected number of pedigrees carrying a post-filtering variant is derived from KCL exomes only, while the observed numbers are from KCL and ICMB pedigrees.)

RGA is performed for both ranking methods using all five (hub-free) networks, with and without jumps being permitted. For the simple count ranking where genes are assigned relatively few distinct ranks, all combinations of α and β thresholds are tested. For the case-control ranking, all thresholds in the range $1 \leq \alpha \leq 1{,}000$ are tested, with $\beta = \alpha$. 10,000 degree-constrained permutations of each network are used to establish whether a significantly large region is identified for any pedigree, with RGA's default standard deviation of 5.0 nodes used for label-shuffling.

8.2.7 Tools Used for Analysis of Results

As described in full in section 7.2.7 in the previous chapter: network diagrams are generated using Cytoscape (Smoot et al. 2011); exploration of existing functional annotation for gene sets is performed using Gene Ontology (GO) biological process term enrichment analysis using WebGestalt (Wang et al. 2013a); and unless otherwise referenced, summaries of individual gene function are obtained from GeneCards (www.genecards.org, Stelzer et al. 2011).

In addition, to assess previous evidence for involvement in CD, results are compared against the list of 300 prioritised genes identified by the IBD GWAS meta-analysis by Jostins *et al*. (Jostins et al. 2012). It is important to recognise that absence from this list does not exclude genes from consideration as causal for familial CD, for several reasons: the meta-analysis focused on sporadic (rather than familial) cases of IBD which we are assuming result from a different disease mechanism, as discussed in section 8.1.1; the list of 300 genes may omit the true disease-relevant genes at some IBD-associated loci due to assumptions made by the underlying prioritisation methods; and the 163 loci reported by Jostins *et al*. do not comprise a definitive list (efforts are ongoing to identify further IBD-associated loci). However, we might expect some overlap between the functional pathways involved in different forms of CD, and thus the presence on this list of a putative familial CD gene could be considered supporting evidence for disease involvement.

Results will also be compared against relevant curated pathways downloaded from the Molecular Signatures Database (MSigDB) on 30th April 2014 (Subramanian et al. 2005).

8.3 Results and Discussion

8.3.1 Post-Filtering Variants in Single Genes

As in the previous chapter, the interpretation of BioGranat-IG results (sets of up to four interacting genes found to contain post-filtering variants for as many CD pedigrees as possible) is aided by a consideration of the individual genes harbouring variants in the highest number of pedigrees; that is, a consideration of what would be revealed by simple intersection filtering.

Table 8.2 lists the genes in which the most CD pedigrees carry a post-filtering sequence variant. It is not practical to look at every gene in this table individually, but several observations will be made here.

ZNF610, which encodes a zinc finger protein, harbours a variant in 14 of the 25 pedigrees. Of the 338 non-CD control exomes, 16 carry a post-filtering variant in this gene, which is below the exclusion threshold of 20, but still suggests the gene may be reasonably tolerant to functional variation. Further, all 14 pedigrees with a *ZNF610* variant are from the 16 pedigrees sequenced at ICMB (none of the pedigrees sequenced at KCL carry a variant) and in all 14 pedigrees the same variant is observed. This allocation between pedigrees is highly unlikely to occur by chance: an empirical test for non-random allocation gave a p-value of $p = 0.0007$ (obtained by 10^6 simulations in which 14 pedigrees were sampled without replacement from the full set of 25, with probabilities proportional to the number of genes in which each pedigree carries a post-filtering variant). This suggests that *ZNF610* is most likely a sequencing artefact seen due to platform differences between the whole exome sequencing performed at ICMB and at KCL. As such, and because all of the non-CD control exomes were sequenced at KCL, the variants in *ZNF610* are likely to be false positives. It would be expected that this gene would be excluded if control exomes from ICMB were available to perform gene-level filtering.

A similar phenomenon is also observed for the next three genes in the list: *MYL12B* has identical variants in 11 of the pedigrees (all of them sequenced at ICMB; empirical allocation p-value $p = 0.0114$), and only 11 of the 338 non-CD controls; *MYO19* has variants in 10 of the pedigrees (all ICMB, $p = 0.0219$; four distinct variants in total but on average 1.8 variants per pedigree) and 10 non-CD controls; and *KCNA6* has identical variants in 8 of the pedigrees (all ICMB; $p = 0.0644$) and 5 non-CD controls. Again, these are assumed to be false positive findings which pass gene-level filtering due to platform differences between whole exome sequencing of cases at ICMB and non-CD controls at KCL.

For this reason, and because these genes will be predisposed to occur in optimal BioGranat-IG subnetworks and RGA regions, the four genes *ZNF610*, *MYL12B*, *MYO19* and

Table 8.2 – Genes with rare non-synonymous variants in the highest number of CD pedigrees
"Count" gives the number of CD pedigrees (of 25 in total) in which each gene contains a variant (after filtering based on frequency and variant consequence as described in section 8.2.2). "Genes after filtering" includes those genes in which fewer than 20 of the 338 non-CD control exomes carry a rare non-synonymous variant; these genes are all taken forward for BioGranat-IG analysis and across-pedigree RGA apart from those in red, which are thought to be false positive findings (see main text). "Genes excluded due to variants in non-CD control exomes" are those genes in which 20 or more of the 338 non-CD control exomes carry a rare non-synonymous variant; none of these genes are taken forward for BioGranat-IG analysis and across-pedigree RGA.

Count	Genes after filtering	Genes excluded due to variants in non-CD control exomes
14 pedigrees	*ZNF610*	
11 pedigrees	*MYL12B*	
10 pedigrees	*MYO19*	
9 pedigrees		*BCAP31*
8 pedigrees	*KCNA6*	
6 pedigrees		*TTN*
5 pedigrees		*TYRP1*
4 pedigrees	*CLEC2B*, *MYO9A*, *NBPF3*, *PCDHB3*	*DNAH1*, *DNAH5*, *FLG*, *MRPL36*, *MUC4*, *NOD2*, *PCNT*, *SSPO*, *SYNE1*
3 pedigrees	*AK2*, *FANK1*, *IK*, *KRTAP2-2*, *PCDHB2*, *PKIB*, *RAD52*, *WDFY3*	17 genes

KCNA6 were removed from the filtered gene lists before performing BioGranat-IG analysis and across-pedigree RGA (see results in sections 8.3.2 and 8.3.4 respectively).

Subsequently in Table 8.2 there are four genes having variants in four pedigrees. For three of the genes all four pedigrees come from the 16 ICMB pedigrees, but this is unremarkable (empirical allocation p-value $p = 0.3185$), while for the other gene (*PCDHB3*) one of the pedigrees was sequenced at KCL. These genes and those with variants in fewer pedigrees will be retained for BioGranat-IG analysis and across-pedigree RGA.

There appears to be no single gene which is a strong candidate to explain the occurrence of CD in a majority of the pedigrees. However, it could be that the gene-level filtering step, which excludes genes in which 20 or more of the 338 non-CD control exomes carry a post-filtering variant, is too restrictive (particularly since the variant-level filtering criteria are more relaxed than the variant-level filtering criteria used for the AOS exomes in chapter 7[*]). We can see that this is not the case by considering the final column of Table 8.2,

[*] Here 1,675 genes are excluded due to the variants they contain in the non-CD control exomes; at filtering level 1 in the previous chapter only 572 genes were excluded due to the variants they contain in the non-AOS control exomes.

which indicates the genes that would be listed were it not for the gene-level filtering step. The first gene in this column is *BCAP31*, and only nine of the pedigrees carry a variant in this gene. So too do 65 of the 338 non-CD control exomes. *TTN* harbours a variant in six of the pedigrees, but is also the gene displaying the most post-filtering variation in controls (248 of 338 non-CD controls carry a variant). These and the other genes in the final column of Table 8.2 are excluded because according to our threshold they tolerate post-filtering variants in the non-CD controls, but it is not true that this gene-level filtering has excluded many otherwise-strong candidates.

Note that *NOD2*, the gene currently thought to explain the most variance in CD disease risk (Jostins et al. 2012), also appears in this table with variants in four pedigrees. This includes two known risk variants, but as discussed earlier we suspect that these pedigrees may carry other rare variants of interest. *NOD2* contains a post-filtering variant in 35 of the non-CD control exomes (which may or may not include variants that have a small effect on CD susceptibility, as noted in section 8.2.1) and is therefore one of the genes excluded by the gene-level filtering step.

We can conclude, therefore, that no single gene shows strong evidence for causality in a majority of pedigrees. Therefore it is reasonable to assume that locus heterogeneity may be present, so that studying the CD pedigrees in a network context is justified.

8.3.2 BioGranat-IG Results

As in chapter 7, most attention is given to the searches in the PINA_d50 network since this is the network of physical interactions with widest genomic coverage (10,375 genes); subsequently KGGSeq-prioritisation will be used to identify any additional subnetworks from the remaining networks that warrant further attention.

8.3.2.1 Results Summary

In total, ten different BioGranat-IG searches were performed using each combination of five different networks (PINA_d50, PINAmin2_d50, CPDBconf95_d50, COXPRES30_d50 and Multinet_d50) and two different search methods (exact triplet and quadruplet searches).

Table 8.3 summarises the findings for each search, giving the number of genes in an optimal subnetwork and the number of CD pedigrees in which an optimal subnetwork harbours a variant. Note that optimal subnetworks are not necessarily unique because several equivalently good subnetworks might be found by a given search (optimality implies that no subnetworks were found to harbour variants in a greater number of CD pedigrees and no smaller subnetworks were found to harbour variants in the same number of CD pedigrees).

Table 8.3 – Summary of optimal subnetworks for familial CD found by BioGranat-IG
Results take the form: *number of CD pedigrees covered (number of genes)*. There are 25 CD pedigrees in total. Note this table gives the properties of optimal subnetworks found by each search but optimal subnetworks are not necessarily unique.

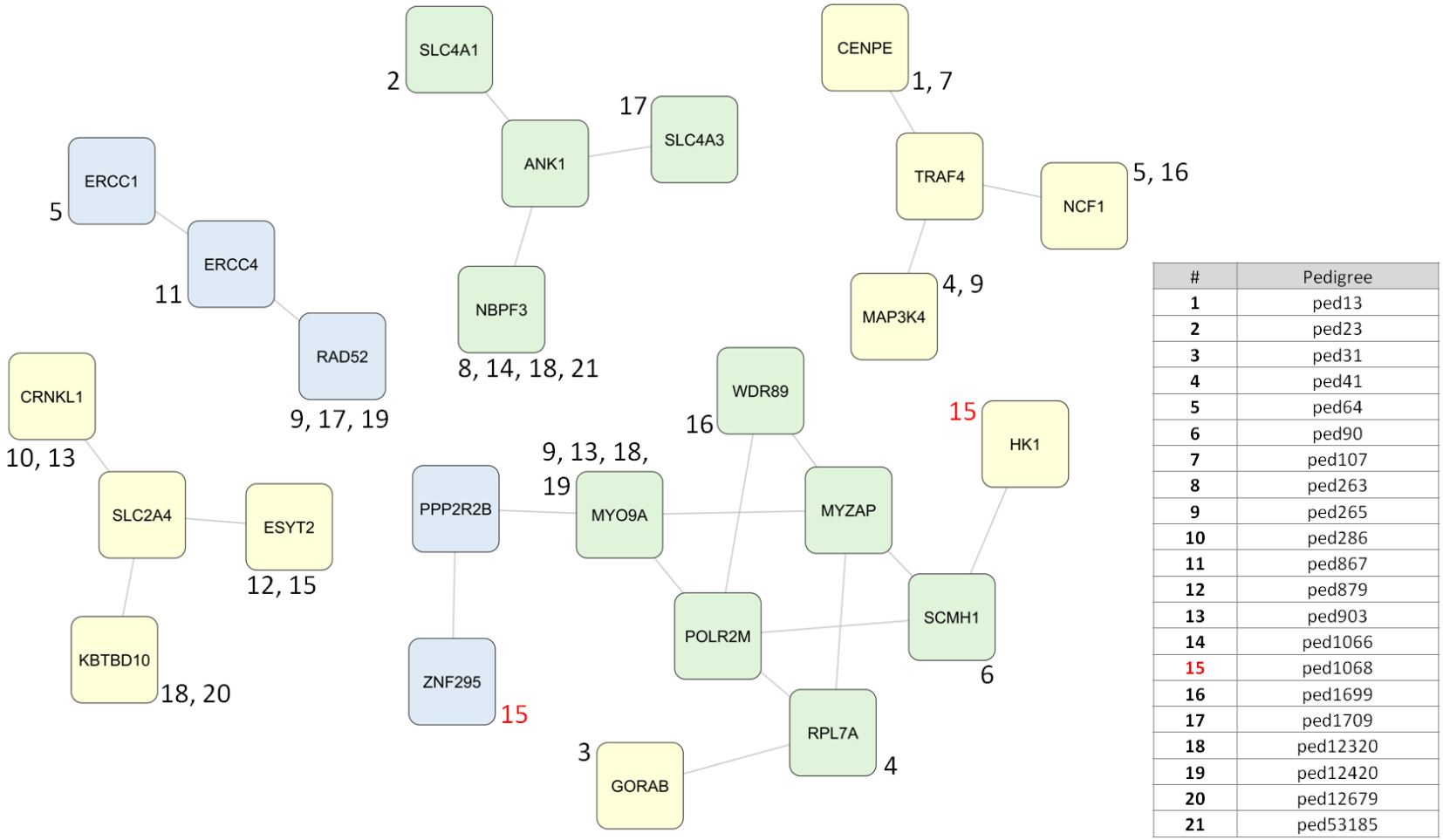
Network	Triplet search	Quadruplet search
PINA_d50	5 (3)	6 (4)
PINAmin2_d50	4 (3)	5 (4)
CPDBcon95_d50	5 (3)	5 (3)
COXPRES30_d50	7 (3)	7 (3)
Multinet_d50	5 (3)	6 (4)

8.3.2.2 PINA_d50 Network

Optimal triplets found by BioGranat-IG in PINA_d50 harbour variants in five of the CD pedigrees. There are ten such subnetworks, which can be merged to form three distinct regions in the network, as depicted in Figure 8.4.

One region comprises a single triplet consisting of the genes *ERCC1*, *ERCC4* and *RAD52*. *ERCC1* and *ERCC4* encode proteins forming a heterodimer that functions in the nucleotide excision repair pathway, and a GO enrichment test suggests that all three genes are involved in "DNA recombination" ($adjP = 6.43\times10^{-5}$) and "double-strand break repair" ($adjP = 1.45\times10^{-5}$). The KGGSeq-prioritisation score is not significant (probability of containing a disease-causing variant = 0.3595, $p = 0.3383$), and none of the genes are included in the list of 300 prioritised genes identified by the CD GWAS meta-analysis by Jostins *et al*. (Jostins et al. 2012). There is therefore little evidence from existing functional annotation or previous association studies that the variants in this subnetwork could cause CD.

A second region comprises two triplets which include the genes *ANK1*, *NBPF3*, *SLC4A1* and *SLC4A3*. In total these genes harbour variants in six of the pedigrees, although no variants occur in *ANK1*, the gene that connects the other three in the network. This region is found primarily because of the gene *NBPF3*, in which the same 6 bp deletion is identified in four of the CD pedigrees. Therefore, whether this region is of interest depends mainly on whether *NBPF3* is likely to have a role in CD. *NBPF3* is a member of the neuroblastoma breakpoint family, a group of recently-duplicated genes containing many tandem repeats of the DUF1220 protein domain (which has unknown function). This raises the possibility that the variants observed are the result of sequencing or alignment errors (inspection of the sequencing reads against the segmental duplication track (Bailey et al. 2001; Bailey et al. 2002) of the UCSC Genome Browser (Kent et al. 2002) suggests that the deletion falls in a duplicated region); to establish whether these variants are genuine, Sanger sequencing could



#	Pedigree
1	ped13
2	ped23
3	ped31
4	ped41
5	ped64
6	ped90
7	ped107
8	ped263
9	ped265
10	ped286
11	ped867
12	ped879
13	ped903
14	ped1066
15	ped1068
16	ped1699
17	ped1709
18	ped12320
19	ped12420
20	ped12679
21	ped53185

Figure 8.4 – Optimal subnetworks found in PINA_d50 using triplet and quadruplet searches for CD pedigrees
Merged regions shown. Genes found in optimal triplets are in blue; quadruplets in yellow; overlap in green. The number(s) next to each node refer to the pedigrees in which they contain variants; coloured numbers indicate exomes with multiple variants in the same merged region.

be employed. In practice this is not recommended because *NBPF3* does not appear to be a strong candidate CD gene. Copy number variations of the DUF1220 domain have been linked with several developmental and neurogenetic diseases, while diseases associated with *NBPF3* itself include schizophrenia and neuroblastoma (www.genecards.org, Stelzer et al. 2011). It is not clear, therefore, how this gene might play a role in CD. Additionally, the KGGSeq-prioritisation score of this network region is not significant (probability of containing a disease-causing variant = 0.3918, p = 0.3213), and none of the genes are included in the list of 300 prioritised genes identified by Jostins *et al*. (Jostins et al. 2012).

Lastly we find a region of eight genes formed by seven overlapping optimal triplets, which in total harbour variants in eight of the pedigrees. Common to all of these triplets is *MYO9A*, in which four of the pedigrees carry a variant; no other gene contains a variant in more than one pedigree, and most of the region is connected via two relatively highly-connected genes, *MYZAP* (with degree 18) and *POLR2M* (with degree 16), which contain no variants themselves. As with the previous region, whether this region could indicate a possible CD disease mechanism or not depends mainly on a single gene, in this case *MYO9A*. This gene encodes a member of the myosin family of motor proteins. Interestingly, mutations in another myosin gene, *MYO9B*, have been associated with IBD and it is hypothesised that they influence intestinal permeability (van Bodegraven et al. 2006). However *MYO9A* itself has not been clearly linked to any disease; it is widely expressed and is thought to be involved in neuron development – which does not readily suggest a role in CD (www.genecards.org, Stelzer et al. 2011). The region as a whole is not enriched for any known function, based on a GO enrichment test (data not shown); the KGGSeq-prioritisation score for the region is not significant (probability of containing a disease-causing variant = 0.4803, p = 0.3344); and none of the genes feature in the list of 300 prioritised genes identified by Jostins *et al*. (Jostins et al. 2012) (although *MYZAP* and *POLR2M* do interact directly with one gene in the list, *CEBPG*).

In summary, based on this preliminary examination none of the regions identified by the optimal triplet search are strongly suggestive of a functional role relevant to CD, and moreover none harbour variants for a majority of the pedigrees.

In the case of the quadruplet search in PINA_d50, we find that optimal subnetworks of four genes harbour variants in six of the CD pedigrees. There are 13 such subnetworks, which can be merged to form four distinct regions in the network; these are also depicted in Figure 8.4.

One region comprises the quadruplet *ANK1-NBPF3-SLC4A1-SLC4A3*, and another is based around the gene *MYO9A*. These were discussed above because they were also identified by the triplet search.

In addition we find a quadruplet comprising the genes *CRNKL1*, *ESYT2*, *KBTBD10* and *SLC2A4*. Of these genes, *SLC2A4* does not contain a variant for any of the pedigrees, but acts to connect together the three other genes in the network. Since this gene is relatively highly connected, with 36 direct neighbours in PINA_d50, this finding is likely to have occurred by chance and not be of relevance to CD. This conclusion is supported by the fact that a GO enrichment test finds no statistically significant functional annotation and the quadruplet does not have a significant KGGSeq-prioritisation score (data not shown).

Finally we see a quadruplet comprising the genes *CENPE*, *MAP3K4*, *NCF1* and *TRAF4*. Again one gene (*TRAF4*) is a connecting gene that contains no variants itself, but has a relatively high degree of 25. The quadruplet's KGGSeq-prioritisation score is far from significant (probability of containing a disease-causing variant = 0.1961, $p = 0.9827$). These factors would suggest that the quadruplet is unlikely to represent a candidate disease pathway. However, a GO enrichment test suggests that *CENPE*, *MAP3K4* and *TRAF4* are involved in "positive regulation of kinase activity"/"regulation of protein phosphorylation" (*adjP* = 0.0059 in each case), which are of potential relevance to CD given the role that phosphorylation signalling plays in the immune system (Cohen 2014). Further, *TRAF4* is a direct neighbour in PINA_d50 of *TNFRSF4*, which encodes a member of the tumour necrosis factor receptor superfamily and has an immune signalling role (and is also annotated with the GO terms "regulation of kinase activity"/"regulation of protein phosphorylation"). *TNFRSF4* was highlighted by Jostins *et al*. as a possible candidate gene tagged by the SNP rs12103, which displayed genome-wide significant association in their IBD GWAS meta-analysis (which combines CD and UC cases), including in their CD-specific test (Jostins et al. 2012). Finally, variants in *NCF1* have been shown to cause chronic granulomatous disease, an early-onset disease which can present with CD-like intestinal inflammation (Uhlig et al. 2014). For these reasons, further study of this subnetwork may be justified.

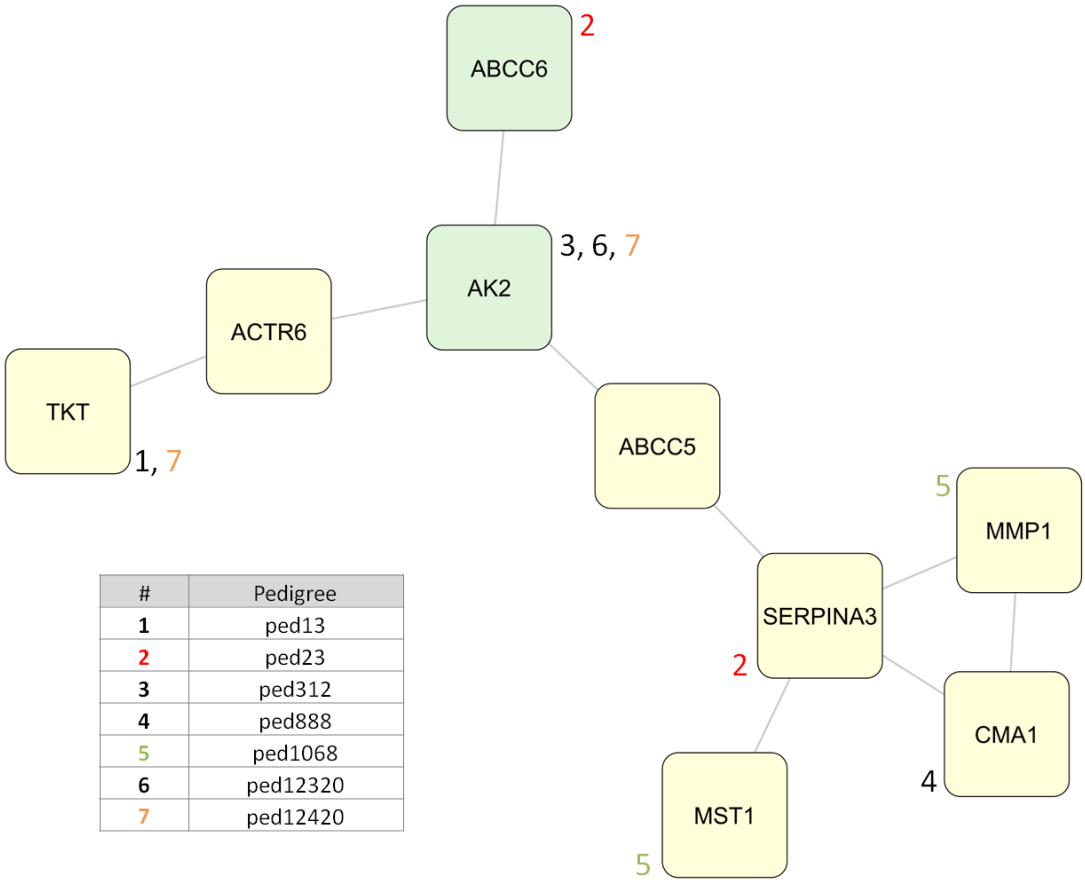
Consideration of near-optimal subnetworks from the triplet and quadruplet searches is also worthwhile because these may implicate different network regions that are enriched for post-filtering variants in CD pedigrees, but for which any given set of three or four connected constituent genes does not quite achieve optimality. Since the BioGranat-IG parameters for near-optimal searches specify size flexibility = 1 and number flexibility = 1, near-optimal subnetworks can contain variants for one pedigree fewer than optimal subnetworks, but can only be up to one gene bigger than the smallest subnetwork that does this. For the triplet search, since an optimal subnetwork comprises three genes with variants in five CD pedigrees, but the smallest subnetworks having variants in four CD pedigrees are in fact single genes (*MYO9A* and *NBPF3* being the only two in PINA_d50), near-optimal

subnetworks are those of up to two genes that contain variants in four CD pedigrees. For the quadruplet search, where optimal subnetworks comprise four genes with variants in six CD pedigrees, near-optimal subnetworks are those of four genes with variants in five CD pedigrees.

For the triplet search, near-optimal subnetworks identify the same regions as were found by the optimal triplet search. In addition we see a region comprising the two genes *ABCC6* (variant in one CD pedigree) and *AK2* (variant in three CD pedigrees) (see Figure 8.5). These genes share no statistically significant enrichment for GO biological process annotation and do not have a significant KGGSeq-prioritisation score (data not shown), and neither features in the list of 300 prioritised genes identified by Jostins *et al*. (Jostins et al. 2012).

However, both genes are part of a region of nine genes found by the near-optimal quadruplet search, shown in Figure 8.5. This region harbours variants in seven of the CD pedigrees, with two variants for pedigrees 23, 1068 and 12420. In the previous chapter, when a region contained multiple variants for the same exome this suggested at least one of the variants were not relevant to the disease process; this was because AOS was not expected to be caused by more than one sequence variant in an affected individual. For familial CD, however, we do not hold this prior belief about the genetic architecture and the fact that these pedigrees carry two variants in a putative disease pathway may constitute increased evidence of involvement. A GO enrichment test finds only one significantly associated biological process term, "extracellular matrix disassembly" (*CMA1* and *MMP1*; *adjP* = 0.0153), and the KGGSeq-prioritisation score is not significant (probability of containing a disease-causing variant = 0.6642, $p = 0.1575$). Note, however, that *MST1* was highlighted by Jostins *et al*. as a possible candidate gene tagged by the SNP rs3197999, which displayed genome-wide significant association in their IBD GWAS meta-analysis (which combines CD and UC cases), including in their CD-specific test (Jostins et al. 2012).

Additionally the near-optimal quadruplet search identifies a region of 132 genes which contain variants for 23 of the 25 CD pedigrees. This region includes the genes from the various regions found by the optimal quadruplet search around the genes *MYO9A*, *TRAF4* and *SLC2A4* (see Figure 8.4), as well as the optimal triplet *ERCC1-ERCC4-RAD52*. Since 132 genes are too many to suggest a closely-related functional pathway underlying mono/oligogenic CD, it is more informative to study the smaller regions derived from the optimal searches, as has been discussed previously.



#	Pedigree
1	ped13
2	ped23
3	ped312
4	ped888
5	ped1068
6	ped12320
7	ped12420

Figure 8.5 – Near-optimal subnetworks found in PINA_d50 using triplet and quadruplet searches for CD pedigrees
Merged region shown. Genes found in near-optimal quadruplets are in yellow; overlap with near-optimal triplets in green. The number(s) next to each node refer to the pedigrees in which they contain variants; coloured numbers indicate exomes with multiple variants in the region. Not shown: regions based around the genes already identified by the optimal searches (and shown in Figure 8.4).

8.3.2.3 Top Prioritised Results in all Networks

None of the results found in PINA_d50 had a significant KGGSeq-prioritisation score, and the same was found to be true of the other four networks. Table 8.4 lists the ten optimal subnetworks (and merged regions of optimal subnetworks) that achieved the smallest KGGSeq-prioritisation p-values across all tests in all networks. Two of these subnetworks (quadruplets containing the gene *MYO9A*) were found in PINA_d50, discussed in the previous section.

The other eight subnetworks were all found in PINAmin2_d50, one of the high-confidence PINs. The most significant KGGSeq-prioritisation score is achieved by an optimal quadruplet comprising the genes *ACTB*, *CCT4*, *NCF1* and *SMARCE1* and harbouring post-filtering variants for five CD pedigrees (probability of containing a disease-causing variant = 0.6527, $p = 0.0668$). Slightly less significant is the merged region of nine

Table 8.4 – Top BioGranat-IG optimal subnetworks by KGGSeq-prioritisation score for CD
No optimal subnetworks or merged regions had a nominally significant KGGSeq-prioritisation score. In lieu of this, the table shows the top ten optimal subnetworks or merged regions by KGGSeq-prioritisation p-value. Also shown are the most significant scoring optimal subnetworks or merged regions found in networks with no optimal subnetwork or merged region in the top ten.

	Network	Search method	Subnetwork or merged region	# Genes	# CD pedigrees	Prob. disease causing	p-value	Genes
1	PINAmin2_d50	Quadruplet	Subnetwork	4	5	0.652719	0.06679	*ACTB*, *CCT4*, *NCF1*, *SMARCE1*
=2	PINA_d50	Quadruplet	Subnetwork	4	6	0.537896	0.07695	*HK1*, *MYO9A*, *MYZAP*, *SCMH1*
=2	PINA_d50	Quadruplet	Subnetwork	4	6	0.537896	0.07695	*HK1*, *MYO9A*, *POLR2M*, *SCMH1*
4	PINAmin2_d50	Quadruplet	Merged region	9	9	0.739091	0.08228	*ACTB*, *ACTG1*, *CCT4*, *CCT8*, *NCF1*, *NCF4*, *PBRM1*, *PRKCB*, *SMARCE1*
5	PINAmin2_d50	Quadruplet	Subnetwork	4	5	0.502803	0.08527	*ITK*, *PLCG1*, *SH3BP2*, *SOS2*
6	PINAmin2_d50	Triplet	Subnetwork	3	4	0.447994	0.10201	*PLCG1*, *SH3BP2*, *SOS2*
7	PINAmin2_d50	Triplet	Subnetwork	3	4	0.441567	0.10633	*ITK*, *PLCG1*, *SOS2*
8	PINAmin2_d50	Quadruplet	Subnetwork	4	5	0.492858	0.13756	*KDR*, *PLCG1*, *SH3BP2*, *SOS2*
9	PINAmin2_d50	Quadruplet	Subnetwork	4	5	0.488271	0.14046	*KDR*, *NRP1*, *PLCG1*, *SOS2*
10	PINAmin2_d50	Quadruplet	Subnetwork	4	5	0.486953	0.14174	*ITK*, *KDR*, *PLCG1*, *SOS2*
31	Multinet_d50	Triplet	Subnetwork	3	5	0.365272	0.28635	*GRINL1A*, *MYO9A*, *SCMH1*
39	CPDBconf95_d50	Triplet	Subnetwork	3	5	0.359492	0.31414	*ERCC1*, *ERCC4*, *RAD52*
80	COXPRES30_d50	Triplet	Subnetwork	3	7	0.377081	0.56055	*PCDHB15*, *PCDHB3*, *PCDHGA1*

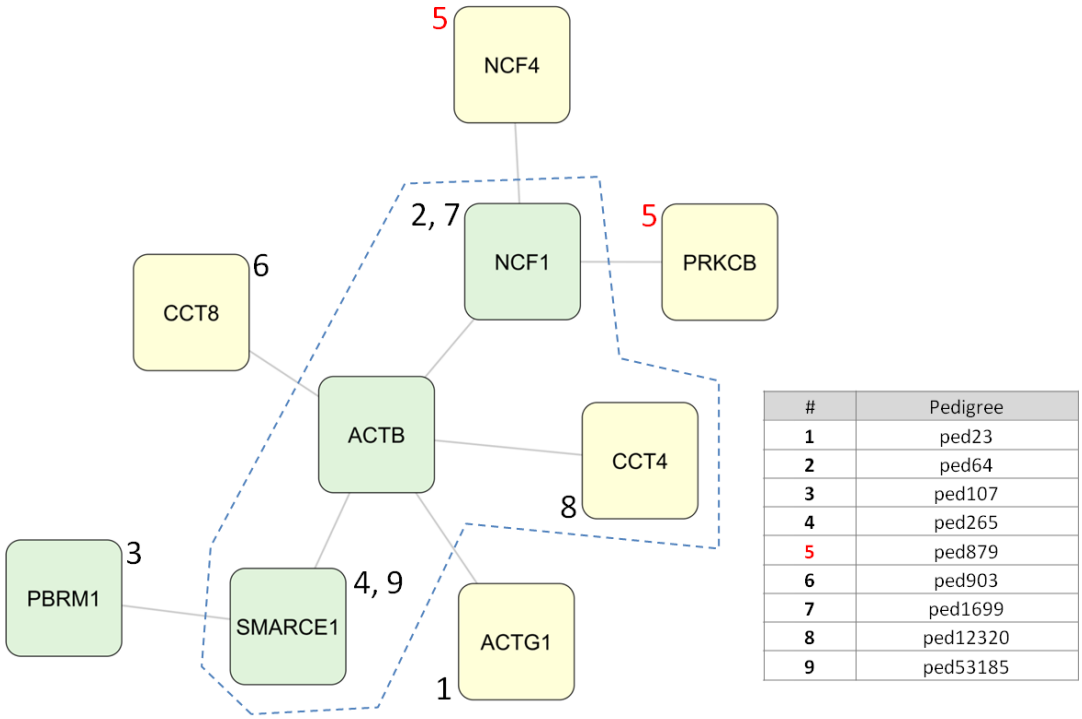
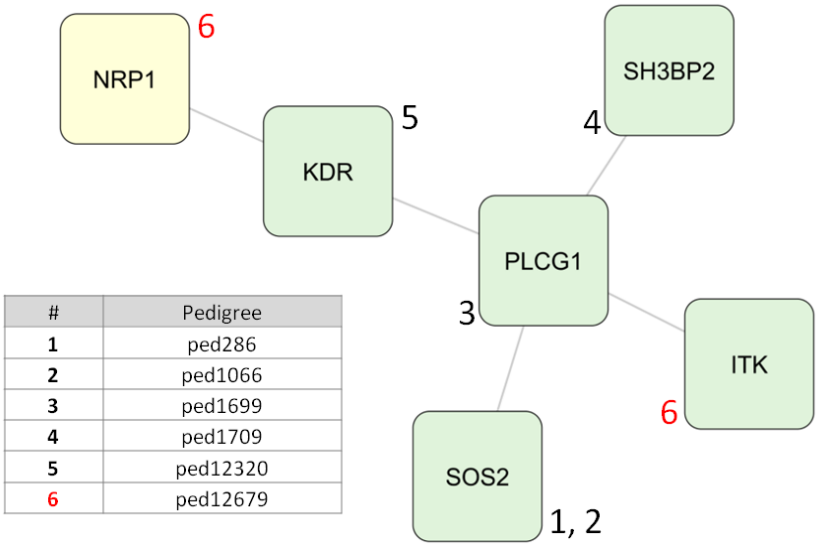


Figure 8.6 – Region of optimal CD quadruplets found in PINAmin2_d50 having most significant KGGSeq-prioritisation scores
Blue dashed region = optimal quadruplet described in row 1 of Table 8.4; all genes = merged region described in row 4. Genes found in quadruplets in PINAmin2_d50 are in yellow; overlap with optimal triplets in green (full results not shown). The number(s) next to each node refer to the pedigrees in which they contain variants; coloured numbers indicate exomes with multiple variants in the region.

genes formed from this and five other optimal quadruplets (probability of containing a disease-causing variant = 0.7391, p = 0.0823) (see Figure 8.6). The near-significant KGGSeq-prioritisation scores achieved by this subnetwork and merged region are largely driven by a novel heterozygous missense SNV that pedigree 12320 carries in *CCT4*. KGGSeq annotates this variant with a probability of 0.5752 of causing a monogenic disease, the fourth highest probability among all post-filtering variants in genes found in PINAmin2_d50. *CCT4* itself encodes a subunit of a molecular chaperone complex involved in protein-folding, which does not immediately present a clear link to the CD phenotype. This is reflected in a GO enrichment test using all nine genes, where the most significantly enriched biological process annotation is for "'de novo' posttranslational protein folding" (*ACTB*, *CCT4* and *CCT8*; *adjP* = 0.0017). Significant enrichment was also observed for "phagosome maturation" (*NCF1* and *NCF4*; *adjP* = 0.0053), which may be of more direct relevance to CD: both of these neutrophil cytosolic factor genes have been shown to cause chronic granulomatous disease, which can cause CD-like intestinal inflammation (Uhlig et al. 2014). Of the nine genes, *PRKCB* is highlighted by Jostins *et al*. as a possible candidate gene tagged by rs7404095, which displayed genome-wide significant association in their



#	Pedigree
1	ped286
2	ped1066
3	ped1699
4	ped1709
5	ped12320
6	ped12679

Figure 8.7 – Additional region of optimal CD triplets and quadruplets found in PINAmin2_d50 and having near-significant KGGSeq-prioritisation scores
Region shown merges optimal subnetworks described in rows 5-10 of Table 8.4. Genes found in quadruplets in PINAmin2_d50 are in yellow; overlap with optimal triplets in green (full results not shown). The number(s) next to each node refer to the pedigrees in which they contain variants; coloured numbers indicate exomes with multiple variants in the region.

IBD GWAS meta-analysis (which combines CD and UC cases) although does not quite reach genome-wide significance in their CD-specific test (Jostins et al. 2012). This gene encodes a kinase involved in many cellular functions; this includes an established immune role – it is an element of Reactome's curated "Activation of NF-κB in B-cells" pathway (obtained via MSigDB).

The remaining six subnetworks having the most significant KGGSeq-prioritisation scores are all optimal triplets or quadruplets in PINAmin2_d50 and each contain both *PLCG1* and *SOS2*. The near-significant prioritisation scores are primarily driven by a novel heterozygous missense SNV that pedigree 1066 carries in *SOS2*. KGGSeq annotates this variant with a probability of 0.3166 of causing a monogenic disease. The six subnetworks form a region of six genes when considered together (see Figure 8.7). A GO enrichment test based on all six genes suggests significant shared function related to "positive regulation of endothelial cell migration" (*KDR*, *NRP1* and *PLCG1*; *adjP* = 0.0007) and "activation of phospholipase C activity" (*ITK* and *PLCG1*; *adjP* = 0.0102). Inspection of relevant curated pathways from MSigDB highlights the presence of *ITK* and *PLCG1* in Reactome's "adaptive immune system" pathway (the latter more specifically involved in "cytokine signalling in the immune system") and *KDR* in KEGG's "Cytokine – cytokine receptor interaction" pathway. These are of note because one of the key molecular mechanisms underlying CD, the IL23 pathway, involves cytokine signalling (Brand 2009). *SOS2* itself encodes a guanine nucleotide exchange factor and is a member of KEGG's "JAK/STAT

signalling pathway". None of the six genes are included in the list of 300 prioritised genes identified by Jostins *et al*. (Jostins et al. 2012) (although *KDR* interacts directly with one gene on the list, *FYN*). Considering these findings together, there is only limited direct evidence that this region could be responsible for CD in six of the pedigrees, but the existing functional annotation of these genes could justify further investigation.

For the remaining networks, Table 8.4 lists the optimal subnetwork or merged region having the most significant KGGSeq-prioritisation score (although none of these are among the ten most significant across all tests).

In Multinet_d50, the network that combines multiple interaction types, the highest-placed subnetwork (ranked 31st) is a triplet that includes the gene *MYO9A* (probability of containing a disease-causing gene = 0.3627, p = 0.2864). Optimal triplets in this network harbour variants for five pedigrees. This triplet was not present in PINA_d50, where optimal triplets also contain variants for five pedigrees, but since *MYO9A* contains a variant in four of the CD pedigrees the same logic as presented in section 8.3.2.2 applies (namely that the relevance of this subnetwork to CD is primarily dependent on the relevance of *MYO9A*).

In CPDBconf95_d50, the high-confidence PIN, the highest-placed subnetwork (ranked 39th) is the triplet *ERCC1-ERCC4-RAD52* (probability of containing a disease-causing gene = 0.3595, p = 0.3141). This triplet was previously found in PINA_d50 (see discussion and Figure 8.4 in section 8.3.2.2). The fact that it is also found in CPDBconf95_d50 suggests that we can be fairly confident in the validity of the interactions between the products of these genes.

Finally in COXPRES30_d50, the co-expression network, the highest-placed subnetwork (ranked 80th) is an optimal triplet comprising the genes *PCDHB3*, *PCDHB15* and *PCDHGA1* and harbouring post-filtering variants in seven CD pedigrees (probability of containing a disease-causing gene = 0.3771, p = 0.5606). Two of these genes are members of the protocadherin beta subfamily. As will be seen in the subsequent results sections, members of this family of genes appear prominently in the results obtained from within-pedigree RGA and across-pedigree RGA. Discussion will therefore be deferred to the next results section.

8.3.3 Within-Pedigree RGA Results

For each network, a binary form of RGA was performed separately for each pedigree to identify connected sets of genes for which the pedigree carries a post-filtering variant (using the relaxed filtering criteria). This was done to suggest possible oligogenic disease mechanisms causing familial CD.

Table 8.5 gives the size of the largest region found for each pedigree in each network. Also given are the number of such largest regions identified, and the p-value generated by RGA that quantifies the degree to which the region found is larger than would be expected under the null hypothesis that regions occur by chance due to the number and degree of genes in the network that contain variants.

In total, 11 significantly large regions were identified. These are summarised in Table 8.6. It may be informative to study these regions directly to assess their likelihood of being relevant to CD. A preliminary analysis of these regions leads to the following observations.

For pedigree 1068 the largest regions identified in PINA_d50 and Multinet_d50 were identical, and the same is true for pedigree 12679 in PINAmin2_d50 and Multinet_d50. Other than this, the only overlap between significant regions was for pedigrees 265 and 888, which both had significantly large regions in COXPRES30_d50 (of nine genes and eight genes; $p = 0.0054$ and 0.0063 respectively; eight genes in common). All of the genes are protocadherin beta genes, one of three clustered families of genes on chromosome 5 encoding cell-adhesion proteins (Chen and Maniatis 2013). This could represent a sequencing artefact due to a high degree of homology among the genes in each cluster. This was investigated by inspection of the whole exome sequencing reads using the UCSC Genome Browser (Kent et al. 2002). For most of the genes in which both pedigrees carried a variant, the variant was identical. The majority of reads appeared to have aligned well to the reference genome (although notably one exome in particular from pedigree 265 had visibly lower quality alignment around several of the called variants). By comparison to the browser's segmental duplication track (Bailey et al. 2001; Bailey et al. 2002), only around half of the variants found appeared to lie in duplicated sequence regions. These observations suggest that at least some of the variants found in these genes may be genuine; one way to confirm this would be to perform Sanger sequencing across the protocadherin beta gene cluster. However, it is unlikely in practice that this would be justified given that these genes appear to have little relevance to CD, being primarily expressed in the nervous system and playing a key role in neurodevelopment (Chen and Maniatis 2013).

For pedigree 90 a region of three genes (comprising *INADL*, *MAPK12* and *RGS3*) was identified in PINA_d50 ($p = 0.0250$). This finding is of potential interest because of *MAPK12*'s role in the innate immune system. Of particular note, given that *NOD2* is the gene currently thought to explain the most variance in CD disease risk (Jostins et al. 2012), *MAPK12* is a member of Reactome's curated "NOD1/2 signalling pathway" (obtained via MSigDB); the MAPK pathway can be activated as a downstream consequence of NOD1 and NOD2 signalling (Windheim et al. 2007).

Table 8.5 – Summary of within-pedigree RGA results for CD

m = size of largest region; *p* = RGA p-value indicating significance of region size; *l* = number of largest regions; l_{LC} = number of largest regions in large component of network. Green shading indicates a significantly large region was found; "-" = no region found.

Pedigree ID	PINA_d50				PINAmin2_d50				CPDBconf95_d50				COXPRES30_d50				Multinet_d50			
	m	*p*	*l*	l_{LC}	*m*	*p*	*l*	l_{LC}	*m*	*p*	*l*	l_{LC}	*m*	*p*	*l*	l_{LC}	*m*	*p*	*l*	l_{LC}
13	2	0.9492	2	2	2	0.7556	1	1	2	0.8499	2	2	3	0.6662	1	1	2	0.8118	2	2
23	5	0.0650	1	1	2	0.9087	2	2	2	0.9793	6	6	5	0.2550	1	1	3	0.7758	1	1
31	2	0.5694	1	1	-	-	-	-	-	-	-	-	2	0.7149	1	1	-	-	-	-
32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
41	2	0.8590	1	1	2	0.5000	1	1	2	0.8549	3	2	2	0.9832	3	2	2	0.5581	1	1
48	2	0.5583	1	1	-	-	-	-	-	-	-	-	2	0.9116	3	3	2	0.3823	1	1
64	2	0.9486	2	2	-	-	-	-	-	-	-	-	2	0.9912	6	6	5	0.0067	1	1
90	3	0.0250	1	1	-	-	-	-	-	-	-	-	3	0.1326	1	1	-	-	-	-
107	2	0.2925	1	1	-	-	-	-	-	-	-	-	3	0.1514	1	1	-	-	-	-
263	2	0.9658	1	1	-	-	-	-	2	0.8915	1	1	4	0.0855	2	2	2	0.9035	1	1
265	4	0.1175	1	1	-	-	-	-	2	0.9123	1	1	9	0.0054	1	1	2	0.9973	9	9
286	2	0.8280	3	3	2	0.3044	2	1	2	0.5341	2	2	3	0.3386	1	1	2	0.8066	3	3
312	2	0.4956	1	1	-	-	-	-	-	-	-	-	2	0.9359	4	4	-	-	-	-
867	2	0.9056	2	2	-	-	-	-	2	0.6845	1	1	2	0.9964	9	9	2	0.8702	1	1
879	3	0.5531	1	1	2	0.8806	2	2	2	0.9400	2	2	4	0.2648	2	2	3	0.4014	1	1
888	2	0.9852	5	5	2	0.9129	3	2	2	0.9543	4	4	8	0.0063	1	1	2	0.9898	4	4
903	2	0.8543	3	3	2	0.6686	2	1	2	0.8088	3	3	3	0.4562	1	1	3	0.3764	1	1
1066	3	0.7860	2	2	2	0.9500	2	2	2	0.9898	4	4	11	0.0025	1	1	2	1.0000	7	7
1068	4	0.0262	1	1	-	-	-	-	2	0.6892	1	1	5	0.0175	1	1	4	0.0057	1	1
1699	2	0.9648	4	4	-	-	-	-	2	0.8730	2	2	3	0.8182	2	2	2	0.9749	5	5
1709	3	0.5528	2	2	2	0.9094	2	2	3	0.3198	2	2	4	0.2096	1	1	2	0.9741	4	4
12320	4	0.3148	1	1	2	0.9858	2	2	2	0.9996	7	7	5	0.5379	1	1	3	0.7101	1	1
12420	4	0.3062	1	1	2	0.9122	2	2	2	0.9879	3	2	7	0.0239	1	1	3	0.4710	1	1
12679	3	0.1773	2	2	4	0.0111	1	1	3	0.1614	1	1	4	0.1100	1	1	4	0.0198	1	1
53185	2	0.3646	1	1	-	-	-	-	2	0.3507	2	2	2	0.9552	1	1	2	0.2808	2	1

Table 8.6 – Significantly large regions found by within-pedigree RGA for CD

Pedigree ID	Network	Region size	P-value	Genes
64	Multinet_d50	5	0.0067	*ALDH1B1*; *ASMT*; *LIPA*; *LPL*; *PTPN4*
90	PINA_d50	3	0.0250	*INADL*; *MAPK12*; *RGS3*
265	COXPRES30_d50	9	0.0054	*PCDHB2*; *PCDHB3*; *PCDHB4*; *PCDHB7*; *PCDHB10*; *PCDHB11*; *PCDHB12*; *PCDHB14*; *PCDHB16*
888	COXPRES30_d50	8	0.0063	*PCDHB2*; *PCDHB3*; *PCDHB4*; *PCDHB7*; *PCDHB10*; *PCDHB11*; *PCDHB12*; *PCDHB16*
1066	COXPRES30_d50	11	0.0025	*DENND4B*; *GON4L*; *INF2*; *MAPK8IP1*; *MYO9B*; *PKD1*; *SHKBP1*; *TAF1C*; *TYK2*; *UBN2*; *WIZ*
1068	PINA_d50	4	0.0262	*GIPC1*; *MMS19*; *PDXDC1*; *TYRP1*
	COXPRES30_d50	5	0.0175	*CLSTN2*; *GFRA1*; *TFF1*; *TFF2*; *TFF3*
	Multinet_d50	4	0.0057	*GIPC1*; *MMS19*; *PDXDC1*; *TYRP1*
12420	COXPRES30_d50	7	0.0239	*ANKRD11*; *GLG1*; *JARID2*; *NUP153*; *SIK3*; *SPG7*; *TCF25*
12679	PINAmin2_d50	4	0.0111	*SMG1*; *TELO2*; *UPF1*; *UPF2*
	Multinet_d50	4	0.0198	*SMG1*; *TELO2*; *UPF1*; *UPF2*

The largest region found is for pedigree 1066, where 11 of the 382 genes containing post-filtering variants are connected in COXPRES30_d50 ($p = 0.0025$). This includes the gene *TYK2*, which was highlighted by Jostins *et al.* as a possible candidate gene tagged by the SNP rs11879191, which displayed genome-wide significant association in their IBD GWAS meta-analysis (which combines CD and UC cases), including in their CD-specific test; the authors also noted that it had been previously implicated in Mendelian susceptibility to mycobacterial disease (Jostins et al. 2012). In addition, *TYK2* encodes a tyrosine kinase which is a member of several curated pathways of potential relevance to CD (obtained via MSigDB), including Reactome's "Cytokine signalling in the immune system" and KEGG's "JAK/STAT signalling pathway". The region also includes *MYO9B*, which has previously been associated with IBD and is thought to influence intestinal permeability (van Bodegraven et al. 2006). As mentioned in the previous chapter, one of the problems with the COXPRES30_d50 network is the functional interpretation of edges: connections indicate correlated expression across a range of gene expression samples but do not directly indicate

Table 8.7 – Clustering of within-pedigree RGA results for CD
"# peds with regions" = number of the 25 CD pedigrees in which within-pedigree regions were identified; "# pairs of peds" = number of ways of choosing two of these pedigrees for comparison; "Average distance" = average taken across all pairs of minimum distance between regions in different pedigrees (first clustering measure); "# overlapping or adjacent" = number of pairs with minimum distance ≤1 (second clustering measure); "Obs." = observed value; "P-value" = probability of same or greater clustering estimated from 10,000 randomly-generated sets of regions. Green shading indicates a significant degree of clustering.

Network	# peds with regions	# pairs of peds	Average distance		# overlapping or adjacent	
			Obs.	P-value	Obs.	P-value
PINA_d50	24	276	1.13	0.0592	16	0.0838
PINAmin2_d50	12	66	1.67	0.5959	2	0.8682
CPDBconf95_d50	18	153	1.00	0.2459	19	0.1761
COXPRES30_d50	24	276	0.83	0.0033	19	0.2071
Multinet_d50	20	190	1.50	0.9018	10	0.7880

involvement in a common process. A GO enrichment test based on all 11 genes finds no significant enrichment for any GO biological process (data not shown).

For pedigree 12420 a region of seven genes was identified in COXPRES30_d50 (p = 0.0239), including *NUP153*. This is of note because of *NUP153*'s role in the immune system, specifically in cytokine signalling (Reactome curated pathway "Cytokine signalling in the immune system" via MSigDB).

In addition to studying the identified regions directly, we also examined all the regions identified (whether significantly large or not) for evidence of clustering in one or more parts of each network. For each network, Table 8.7 gives the observed value for the two measures of clustering used: the average pairwise distance between the largest regions found for different pedigrees, and the number of pairs of pedigree for which overlapping or adjacent regions were found. Also shown are the p-values derived from 10,000 randomly-generated sets of regions, quantifying the degree to which the observed degree of clustering does not occur by chance.

The networks PINA_d50, PINAmin2_d50, CPDBconf95_d50 and Multinet_d50 do not show evidence for clustering of within-pedigree RGA regions using either measure. For the co-expression network COXPRES30_d50, the average distance between largest regions (for the 24 pedigrees for which regions were found) is 0.83, and this value is significantly small (p = 0.0033). Although the number of pedigrees having overlapping or adjacent largest regions was not significant (19 of 276 pairs of pedigrees; p = 0.2071), it is informative to consider the pedigrees for which overlapping regions were found, as these will contribute to the significantly small average distance.

Overlapping regions were found for five pairs of pedigrees, illustrated in Figure 8.8. One of the pairs comprises pedigrees 888 and 1066, whose largest regions are made up of protocadherin beta genes and were discussed earlier. Several of the pedigrees illustrated in the figure had multiple largest regions of only two genes, and since a region of two genes is not significantly large for any pedigree (see Table 8.5) these are unlikely to be of interest in general. One possible exception is the two-gene region consisting of *SULT1A1* and *SULT1A2*, both of which contain post-filtering variants in pedigrees 41 and 48. These genes encode two members of a family of phenol sulphotransferase enzymes, which have a role in chemical metabolism (Gamage et al. 2006). Both genes were highlighted by Jostins *et al*. as possible candidate genes tagged by the same SNP (rs26528), which displayed genome-wide significant association in their IBD GWAS meta-analysis (which combines CD and UC cases), including in their CD-specific test (Jostins et al. 2012).

There is also a connected set of twelve genes in which four pedigrees carry variants. *BRD2* contains a post-filtering variant along with *BCR* and *SMARCA4* for pedigree 107, and with *BPTF*, *EP300*, *MLL5* and *USP42* for pedigree 12320. In addition, *RANBP2* contains a variant along with *EIF3A* for pedigree 31, and with *ATXN7*, *SRSF4* and *USP47* for pedigree 64. Both *EP300* and *RANBP2* have known immune system roles, with *EP300* an element of KEGG's curated "JAK/STAT signalling pathway" and *RANBP2* present in Reactome's "cytokine signalling in the immune system" pathway (both obtained via MSigDB). A GO enrichment test based on all twelve genes finds the region enriched for genes involved in several biological processes: "chromatin modification" (six genes – including five of the seven containing variants for pedigrees 107 and 12320; $adjP = 0.0003$); "protein deubiquitination" (three genes; $adjP = 0.0019$), and "positive regulation by host of viral transcription" (two genes; $adjP = 0.0023$).

As previously discussed, one of the main limitations of the COXPRES30_d50 network is the difficulty in determining the existence and nature of functional relationships underlying the network edges. However, given that we find some significantly large regions and evidence of clustering, together with previous evidence for statistical association with IBD and an immune system role, these regions may justify further study.

8.3.4 Across-Pedigree RGA Results

Network regions that are enriched for post-filtering variants (allowing any number per pedigree) were identified by across-pedigree RGA using two gene-ranking methods: simple count ranking and case-control ranking.

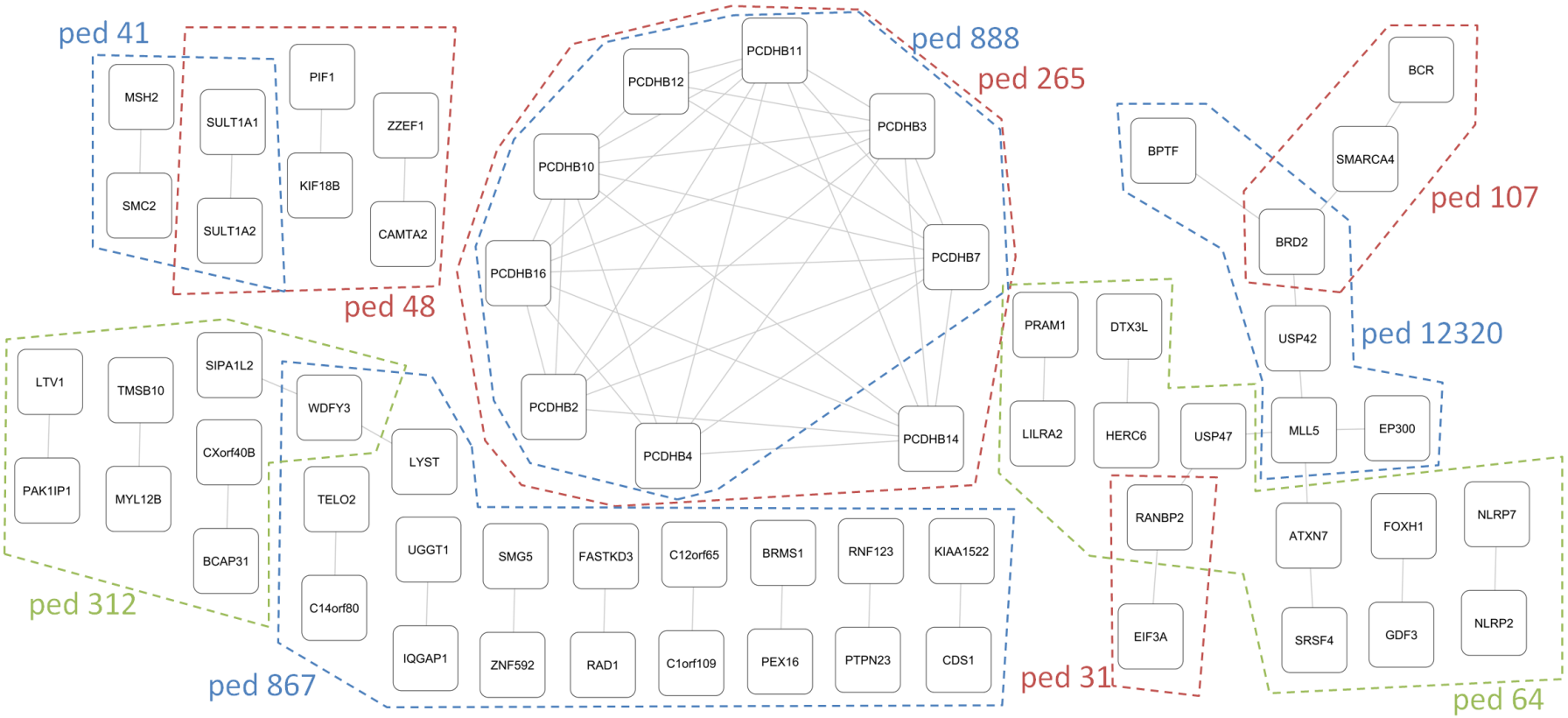


Figure 8.8 – Overlapping within-pedigree RGA regions for CD in the large component of COXPRES30_d50
Dashed boxes indicate largest regions found in each pedigree; there are five overlaps between different pedigrees in total. Note in several pedigrees there were multiple disconnected largest regions found (grouped together in the figure). Colours are used only for ease of interpretation and have no additional meaning.

8.3.4.1 Simple Count Ranking

With genes ranked according to a simple count, RGA was performed using all possible α and β thresholds. These thresholds can be thought of in terms of the number of pedigrees carrying a post-filtering variant in each gene. For example, when α and β correspond to three and two pedigrees respectively, this means that regions are seeded by genes in which three or more pedigrees carry a variant, and are expanded to include member genes in which two or more carry a variant.

A summary of results for all combinations of α and β is given in Table 8.8. As well as the sizes of the largest regions found, the table also gives the empirical p-value that measures the significance of the observed region size (number of seed or member nodes it contains; see chapter 6, section 6.2) against regions found in 10,000 permuted networks. Regions of nominally significant size ($p < 0.05$) were found in only two of the networks.

In the COXPRES30_d50 network, when RGA was performed without jumps being permitted, the most significant region was identified for α corresponding to genes containing variants for four pedigrees and β corresponding to two pedigrees ($p < 10^{-4}$). This region comprises seven of the protocadherin beta genes that were previously discussed in section 8.3.3 above, and will therefore not be discussed further here.

In the CPDBconf95_d50 high-confidence PIN, when RGA was performed with jumps being permitted, the most significant region was identified for α corresponding to genes containing variants for three pedigrees and β corresponding to one pedigree. This region comprised 827 genes in total, with 354 being seed or member nodes for the region ($p = 0.0013$). The fact that the network is significantly large may be of broad interest, but in practice it highlights too many genes to suggest a closely-related functional pathway underlying mono/oligogenic CD, and should therefore not be prioritised for further study.

In two additional networks, regions were found having empirical p-values that were close to being nominally significant ($p < 0.1$), suggesting weak evidence of non-random enrichment for post-filtering variants in CD case pedigrees (illustrated in Figure 8.9).

In the PINA_d50 network, when jumps were permitted the most significant region was identified for $\alpha = \beta$ corresponding to genes containing variants for two pedigrees. The region comprised twelve genes in total, with seven being seed nodes for the region (Figure 8.9a; $p = 0.0750$). The twelve genes include two sets of four genes that were previously identified by BioGranat-IG as optimal quadruplets: *CENPE*, *MAP3K4*, *NCF1* and *TRAF4*; and *CRNKL1*, *ESYT2*, *KBTBD10* and *SLC2A4* (discussed in section 8.3.2.2; see Figure 8.4). Further, all twelve genes feature in near-optimal BioGranat-IG quadruplets. Of the four additional genes, only *TELO2* is a seed gene (containing variants for two of the pedigrees).

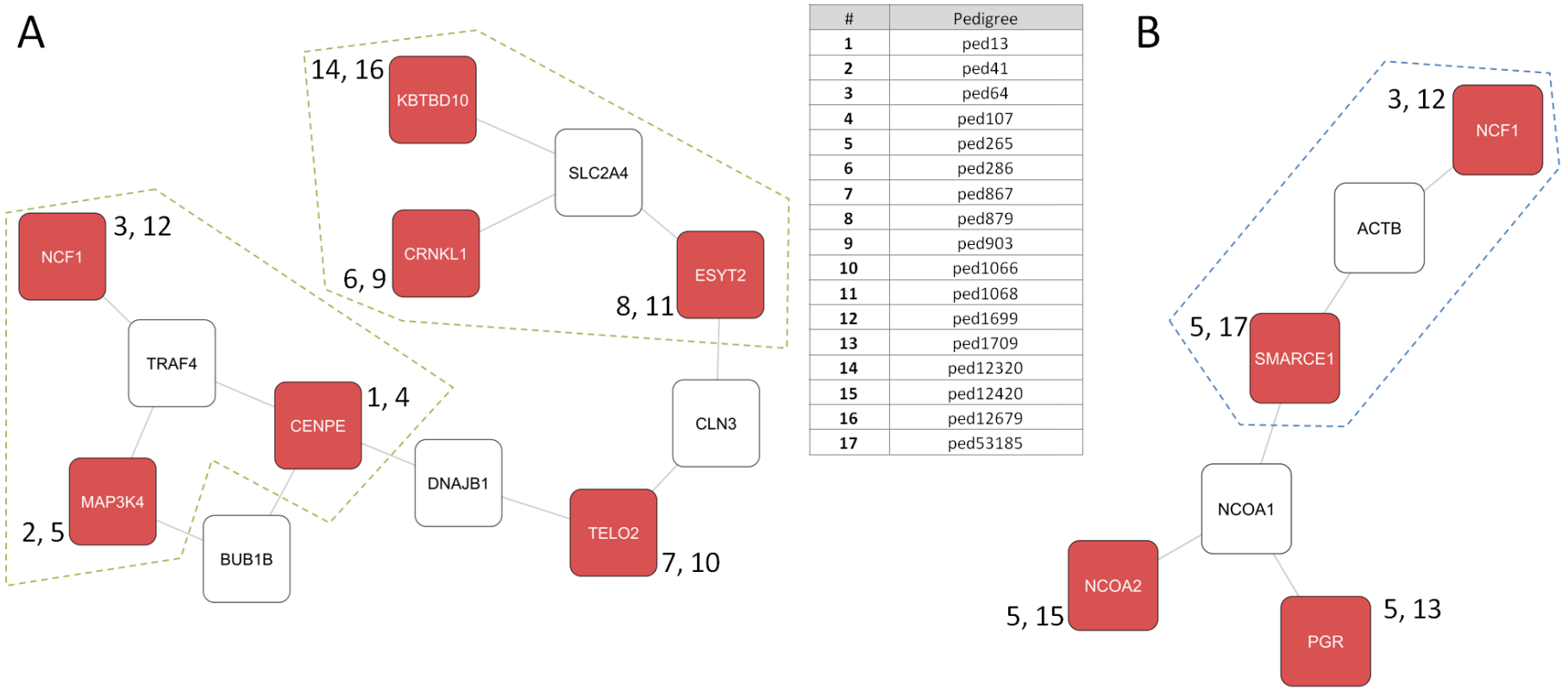
Table 8.8 – Across-pedigree RGA results for CD using simple count ranking
See next page for full table legend.

Network	# pedigrees		No jumps		With jumps		
	α	*β*	Region size	p-value	Region size	Seed/ member nodes	p-value
PINA_d50	1	1	33	0.3667	1586	677	0.1944
	2	1	33	0.3449	1586	677	0.1944
	3	1	4	0.5325	1586	677	0.1944
	4	1	1	1.0000	1586	677	0.1917
	2	2	1	1.0000	12	7	0.0750
	3	2	1	1.0000	3	2	0.7349
	4	2	1	1.0000	1	1	1.0000
	3-4	3-4	1	1.0000	1	1	1.0000
PINAmin2_d50	1	1	10	0.3249	372	200	0.2443
	2	1	10	0.2055	372	200	0.2443
	3	1	1	1.0000	372	200	0.2041
	2	2	1	1.0000	6	4	0.0759
	3	2-3	1	1.0000	1	1	1.0000
	4	1-4	0	1.0000	0	0	1.0000
CPDBconf95_d50	1	1	19	0.3121	827	354	0.0013
	2	1	19	0.2713	827	354	0.0013
	3	1	3	0.3907	827	354	0.0013
	2	2	1	1.0000	5	3	0.7518
	3	2-3	1	1.0000	1	1	1.0000
	4	1-4	0	1.0000	0	0	1.0000
COXPRES30_d50	1	1	48	0.9144	3998	1182	0.2161
	2	1	48	0.9137	3998	1182	0.2161
	3	1	9	0.9470	3998	1182	0.2161
	4	1	9	0.6110	3998	1182	0.2161
	2	2	7	0.0002	47	25	0.2089
	3	2	7	< 0.0001	47	25	0.2055
	4	2	7	< 0.0001	14	7	0.5979
	3	3	2	0.0360	2	2	0.3403
	4	3	2	0.0237	2	2	0.2193
	4	4	1	1.0000	1	1	1.0000
Multinet_d50	1	1	26	0.2817	1138	512	0.6617
	2	1	26	0.2614	1138	512	0.6617
	3	1	5	0.2076	1138	512	0.6602
	4	1	1	1.0000	1138	512	0.5663
	2	2	1	1.0000	8	3	0.7807
	3	2	1	1.0000	7	4	0.1129
	4	2	1	1.0000	1	1	1.0000
	3-4	3-4	1	1.0000	1	1	1.0000

Table 8.8 – Across-pedigree RGA results for CD using simple count ranking (previous page)
"# pedigrees" = definition of the RGA threshold parameters α and β in each test according to the number of pedigrees in which genes contain post-filtering variants; "No jumps" = Results of tests in which RGA does not permit region expansion to include jumps of up to one non-seed/member node; "With jumps" = Results of tests in which RGA allows such jumps; "Region size" = number of nodes in a region; "Seed/member nodes" = number of nodes with rank $\leq\beta$ (here β corresponds to a number of pedigrees in which genes contain post-filtering variants, so seed/member nodes are genes containing variants in this many pedigrees or greater); "p-value" = standard p-value produced by RGA, calculated as the proportion of 10,000 degree-constrained permuted networks in which regions with greater or equal number of seed/member nodes are identified. Green shading indicates nominally significant p-value of < 0.05; yellow shading indicates < 0.1.

None of the twelve genes feature in the list of 300 prioritised genes identified by Jostins *et al*. (Jostins et al. 2012), although *TRAF4* and *SLC2A4* directly interact with genes from the list in PINA_d50. A GO enrichment test based on all twelve genes found no significantly enriched biological process annotation (data not shown). As we found in section 8.3.2.2., the quadruplet *CENPE-MAP3K4-NCF1-TRAF4* alone does show enrichment for a potentially relevant process, suggesting that these four genes should be prioritised for further study (although conversely, as with many of the regions we find, it could be argued that the additional genes in the region could warrant further attention for the purpose of *de novo* pathway discovery).

In the PINAmin2_d50 high-confidence PIN, when jumps were permitted the most significant region was again identified for $\alpha = \beta$ corresponding to genes containing variants for two pedigrees. The region comprised six genes in total, with four being seed nodes for the region (Figure 8.9b; p = 0.0759). One gene (*NCF1*) was also found in the PINA_d50 region discussed in the previous paragraph, and three of the genes (*ACTB*, *NCF1* and *SMARCE1*) were previously identified by BioGranat-IG as an optimal triplet and as part of the optimal quadruplet with the most significant KGGSeq-prioritisation score (see Table 8.4 and Figure 8.6). Of the three additional genes, two (*NCOA2* and *PGR*) are seed genes, each containing post-filtering variants for two of the pedigrees. Pedigree 265 carries a variant in three of the six genes. None of the six genes feature in the list of 300 prioritised genes identified by Jostins *et al*. (Jostins et al. 2012), although there are several direct interactions in PINAmin2_d50 connecting *NCF1*, *NCOA1* and *NCOA2* with genes on the list. To identify shared function among the genes in the region a GO enrichment test was performed. The most enriched terms, including "intracellular steroid hormone receptor signalling pathway" (3 genes; *adjP* = 0.0086) and "cerebellum development" (2 genes; *adjP* = 0.0243), were all based on the function of the two nuclear receptor coactivator genes *NCOA1* and *NCOA2* (only the first of which contains post-filtering variants for any of the CD pedigrees). The nuclear receptor coactivators (or steroid receptor coactivators) are involved in the regulation



#	Pedigree
1	ped13
2	ped41
3	ped64
4	ped107
5	ped265
6	ped286
7	ped867
8	ped879
9	ped903
10	ped1066
11	ped1068
12	ped1699
13	ped1709
14	ped12320
15	ped12420
16	ped12679
17	ped53185

Figure 8.9 – Near-significant regions for CD identified by across-pedigree RGA using simple count ranking
In both subnetworks: red nodes correspond to seed genes; numbers next to each node refer to the pedigrees in which they contain variants. (A) Region of 12 genes identified in PINA_d50 with $\alpha = \beta$ corresponding to post-filtering variants in two pedigrees. Green dashed regions were previously found as optimal quadruplets using BioGranat-IG (shown in Figure 8.4). (B) Region of six genes identified in PINAmin2_d50 with $\alpha = \beta$ corresponding to post-filtering variants in two pedigrees. Blue dashed region was previously found as an optimal triplet and part of an optimal quadruplet using BioGranat-IG (cf. Figure 8.6).

Table 8.9 – Most significant CD regions for across-pedigree RGA using case-control ranking
"No jumps" = results of tests in which RGA does not permit region expansion to include jumps of up to one non-seed node (node with rank >α); "With jumps" = results of tests in which RGA allows such jumps; "Lowest p-value" = lowest empirical RGA p-value observed across tests at each threshold α in $1 \leq \alpha \leq 1{,}000$; "smallest α" = smallest α to have achieved this p-value; "Region size" = number of nodes in largest region at this α threshold; "Seed nodes" = number of nodes with rank $\leq\alpha$ in this region. Green shading indicates nominally significant p-value of < 0.05; yellow shading indicates < 0.1.

Network	No jumps			With jumps			
	Lowest p-value	**Smallest α**	**Region size**	**Lowest p-value**	**Smallest α**	**Region size**	**Seed nodes**
PINA_d50	0.0037	23	3	0.0057	106	31	15
PINAmin2_d50	0.3940	78	2	0.1017	862	171	88
CPDBconf95_d50	0.6197	265	3	0.2221	971	419	185
COXPRES30_d50	0.2139	207	4	0.1024	487	1065	347
Multinet_d50	0.0596	862	12	0.5243	218	18	10

of transcription and consequently influence a wide range of functional systems (York and O'Malley 2010), although there is no established link to IBD.

8.3.4.2 Case-Control Ranking

With genes ranked according to the degree to which the number of pedigrees carrying a post-filtering variant exceeded the expected number (based on a null distribution derived from the 338 non-CD control exomes), RGA was performed using all thresholds in the range $1 \leq \alpha \leq 1{,}000$, with $\beta = \alpha$.

Figure 8.10 shows the size and significance of the largest regions found at each α level for each network, and a summary of the most significant regions identified in each network is given in Table 8.9.

The only network in which region sizes achieved nominal significance ($p < 0.05$) was PINA_d50. When RGA was performed using this network with jumps disallowed, the lowest p-value found was 0.0037, and the first threshold for which this value was observed was $\alpha = 23$. The resulting region comprised three genes: *PPIB*, *SURF4* and *TMBIM4* (see Figure 8.11). All three genes contained a post-filtering variant for one of the 25 pedigrees (pedigrees 1066, 879 and 1709 respectively), but no post-filtering variants in any of the 338 non-CD control exomes. It is worth noting that whole exome sequencing for all three of these pedigrees was performed at ICMB, while all non-CD control exomes were sequenced at KCL, so it is possible that this finding represents technical differences between sequencing platforms (but note that only 27 of the 43 genes for which variants were found in case pedigrees but not in non-CD control exomes have variants exclusively in ICMB

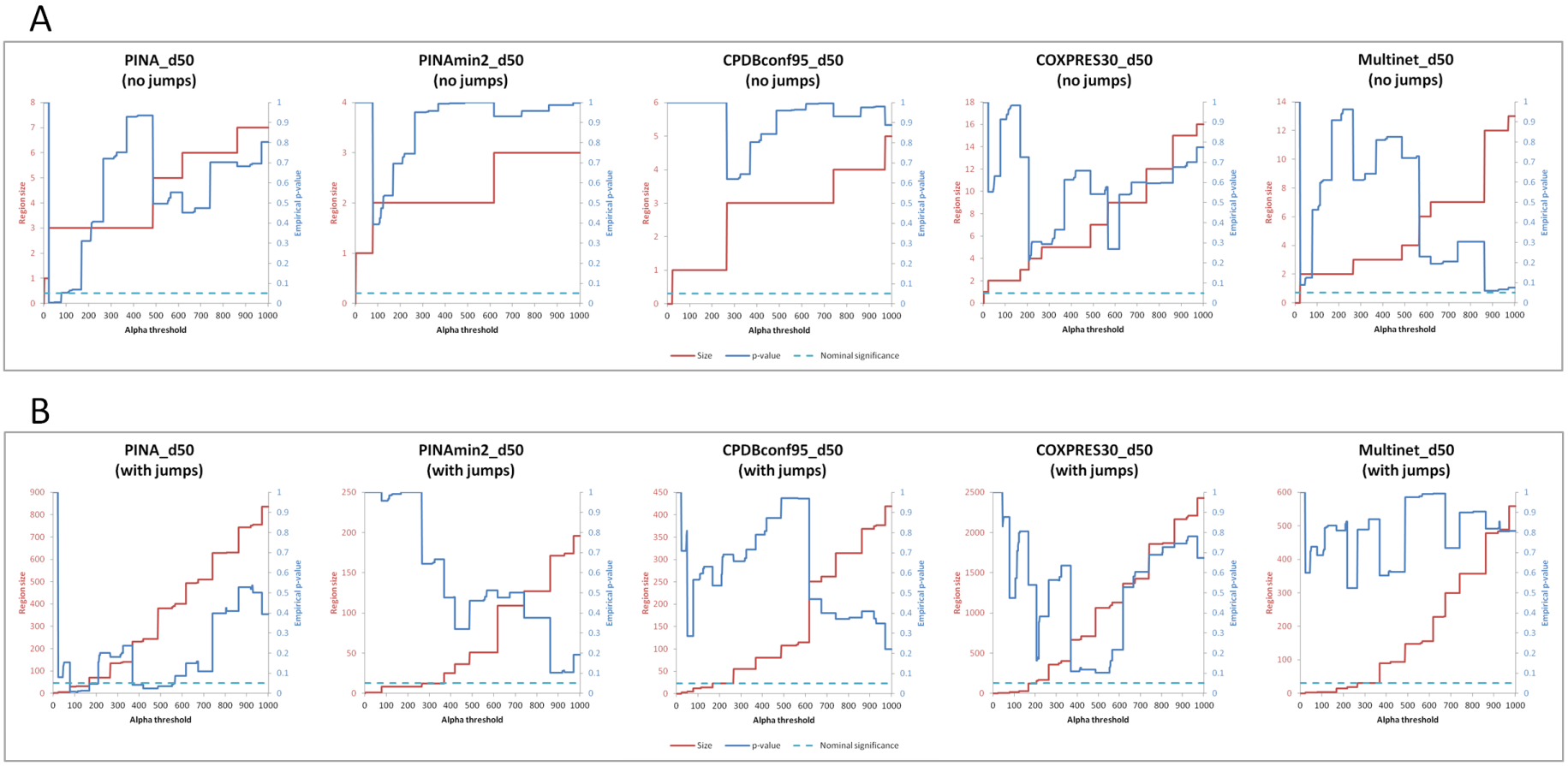


Figure 8.10 – Across-pedigree RGA results for CD using case-control ranking
Each plot gives results for a different network. For each alpha value, size of network region and empirical p-value are plotted. The nominal significance threshold is plotted at 0.05. This output is characteristic of RGA results: the region size is an increasing step function; at each alpha value where there is an increase in region size, the empirical p-value falls; between such alpha values the empirical p-value steadily increases. (A) results when jumps are disallowed; (B) results when jumps are permitted.

#	Pedigree
1	ped23
2	ped31
3	ped41
4	ped64
5	ped263
6	ped265
7	ped286
8	ped312
9	ped879
10	ped888
11	ped1066
12	ped1699
13	ped1709
14	ped12320
15	ped12420
16	ped12679

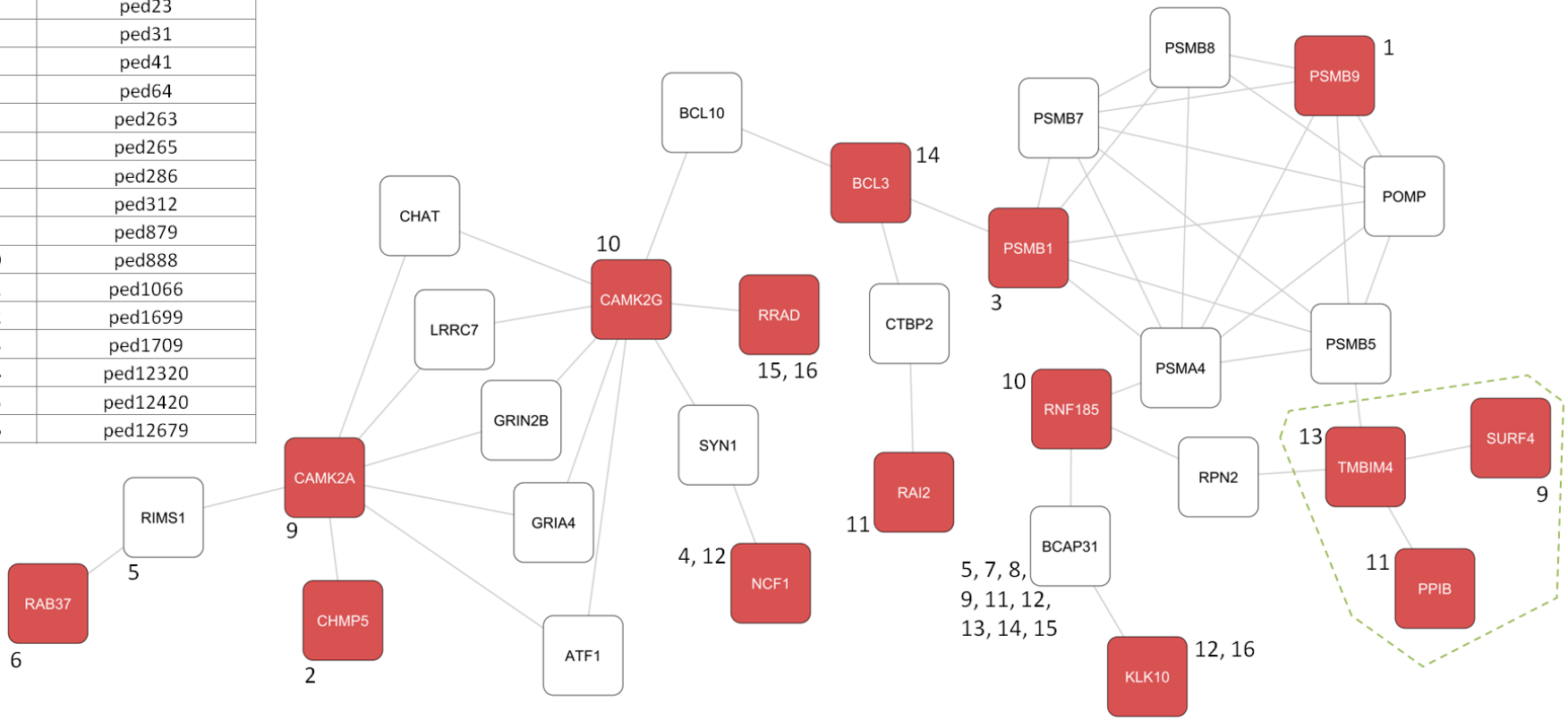


Figure 8.11 – Significant regions for CD identified in PINA_d50 by across-pedigree RGA using case-control ranking
Red nodes correspond to seed genes; numbers next to each node refer to the pedigrees in which they contain variants. Whole region of 31 genes was identified at $\alpha = 106$ when jumps were permitted; three genes within green dashed region were also identified at $\alpha = 23$ when jumps were not permitted.

pedigrees (62.8%); since 16 of the 25 pedigrees were sequenced at ICMB (64.0%) this does not suggest a bias due to platform differences). Since only three pedigrees carry a variant in this region, these three genes are best considered as a part of the region to be discussed in the following paragraph.

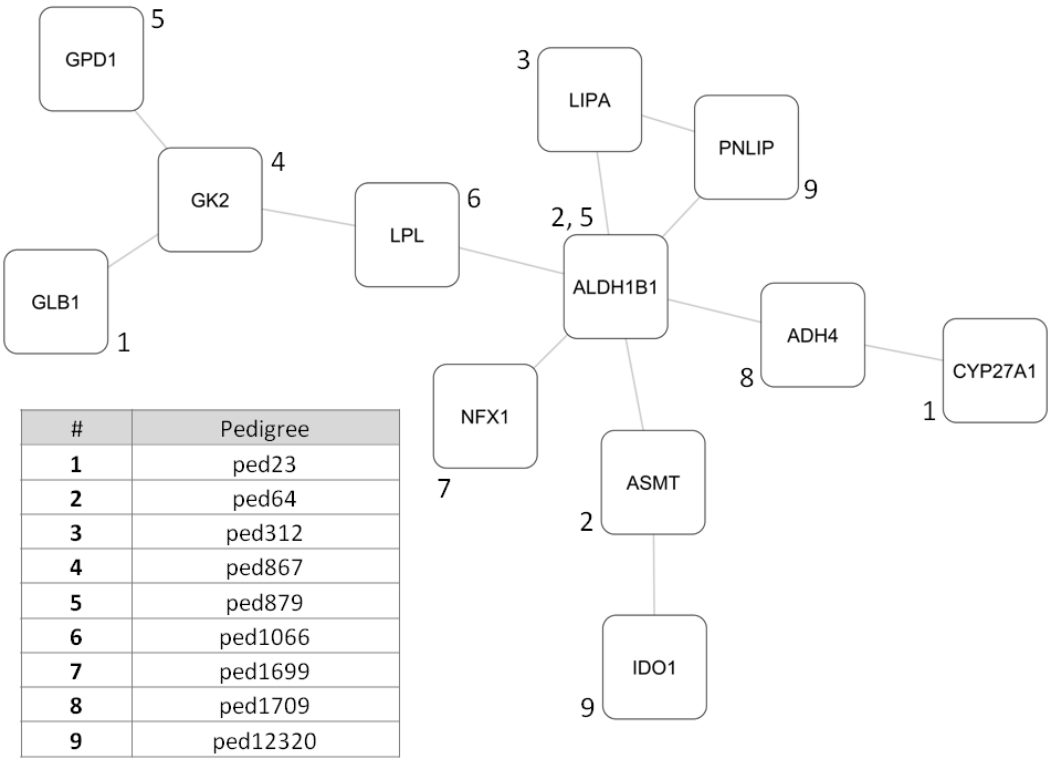
When RGA was performed using PINA_d50 with jumps permitted, the lowest p-value found was 0.0057, and the first threshold for which this value was observed was $\alpha = 106$. The resulting region comprised 31 genes, of which 15 are seed nodes (including *PPIB*, *SURF4* and *TMBIM4* which were found when jumps were not permitted) (see Figure 8.11). This region is intriguing due to the existing functional annotation for several of the genes. It includes a highly-connected module of six proteasome subunit genes, of which *PSMB1* and *PSMB9* are seed nodes (*PSMB1* has a rank of 23 with a post-filtering variant in one pedigree and none in the 338 non-CD control exomes; *PSMB9* has a rank of 77.5 with a variant in one CD pedigree and one non-CD control exome). The other subunit genes do not contain post-filtering variants in any CD pedigree. The proteasome has an important role in the adaptive immune system: it digests cytoplasmic proteins, including antigens which can be presented to the immune system by major histocompatibility complex (MHC) class I molecules (van Kasteren et al. 2014). Thus the most enriched GO biological process term based on all 31 genes in the region is "antigen processing and presentation of peptide antigen via MHC class I", with which all six proteasome genes, plus *BCAP31* and *NCF1*, are annotated ($adjP = 1.07\times10^{-8}$). The six proteasome genes plus *BCL10* are elements of Reactome's curated "activation of NF-κB in B-cells" pathway (obtained via MSigDB). *ATF1* is also present in Reactome's "activated TLR4 signalling" pathway, an innate immune process that responds to bacterial infection. One of the genes, *NCF1*, was found in an optimal BioGranat-IG quadruplet (see Figure 8.4), and four others were found along with *NCF1* in a merged region of 132 genes formed from near-optimal quadruplets (discussed in section 8.3.2.2). As described earlier, we also know that mutations in *NCF1* can result in an early-onset CD-like phenotype (Uhlig et al. 2014). None of the genes feature in the list of 300 prioritised genes identified by Jostins *et al*. (Jostins et al. 2012), although seven have direct interactions in PINA_d50 with genes on this list.

Inspection of the network diagram in Figure 8.11 leads to two further observations. Firstly, note that *BCAP31* is included in the region and contains post-filtering variants in nine genes (as we saw in Table 8.2; recall that for RGA using case-control ranking we have not excluded genes with 20 or more post-filtering variants in the non-CD control exomes). Nevertheless, *BCAP31* is not a seed gene and has been incorporated into the region as a jump. It has a rank of 213 due to the 65 non-CD control exomes in which it contains a variant. Likewise *RIMS1* is included as a jumped node despite containing a variant in one

pedigree. Secondly, of the six proteasome subunit genes, four are included as jumps; only one pedigree actually carries a variant in each of the other two genes. This could have biased the functional enrichment tests described above, and it is perhaps more prudent to examine only genes containing variants (as these are the only ones for which we have direct evidence that could suggest perturbed function in our pedigrees). Repeating the GO enrichment test using only these genes (the 15 seed nodes plus *BCAP31* and *RIMS1*) results in the same biological process ("antigen processing and presentation of peptide antigen via MHC class I") being most enriched (*BCAP31*, *NCF1*, *PSMB1* and *PSMB9*; *adjP* = 0.0017), with "G1/S transition of mitotic cell cycle" also significantly enriched (*CAMK2A*, *CAMK2G*, *PSMB1* and *PSMB9*; *adjP* = 0.0063).

On balance, this region's significant RGA p-value and prior functional annotation make it a potential candidate for further study. Initially this could comprise testing for rare sequence variants in these genes in an additional cohort of familial CD samples, but subsequent laboratory-based experiments (using appropriate cell lines or animal models) would be needed to show that these variants could be linked to the CD phenotype for any of the 16 pedigrees in which the region contains a variant.

The only other network in which a near-significant region ($p < 0.1$) was identified was Multinet_d50. When jumps were not permitted the lowest p-value found was 0.0596, corresponding to $\alpha = 862$ and a region of twelve genes (see Figure 8.12). To give an idea of the enrichment implied by this alpha threshold, the lowest-ranked genes in the region (each having the same rank of 861.5) were *GPD1*, *LPL* and *NFX1*. These genes contain post-filtering variants in one CD pedigree each, and in eight of the 338 non-CD control exomes. Four of the genes (*ALDH1B1*, *ASMT*, *LIPA* and *LPL*) were previously found in a significantly large region found by within-pedigree RGA (see Table 8.6; all four genes and a direct neighbour contain a variant in pedigree 64, but note that relaxed variant-filtering criteria were used in that analysis). None of the genes were previously found in optimal BioGranat-IG triplets or quadruplets, although nine of the twelve genes were found in near-optimal subnetworks (results not shown). None of the twelve genes, and no genes with which they directly interact, feature in the list of 300 prioritised genes identified by Jostins *et al*. (Jostins et al. 2012). The region appears to be broadly implicated in metabolism, the most significantly enriched GO biological process term being "small molecule metabolic process" (11 genes; $adjP = 5.83\times10^{-5}$) with more specific annotation including "alcohol metabolic process" (5 genes; *adjP* = 0.0005) and "lipid metabolic process" (7 genes; *adjP* = 0.0005). Although this may be of broad interest (although diet has not been shown to affect disease risk, there appears to be a complex relationship between nutrition and inflammation in IBD (O'Sullivan and O'Morain 2006)), the non-significant region size and lack of statistical



#	Pedigree
1	ped23
2	ped64
3	ped312
4	ped867
5	ped879
6	ped1066
7	ped1699
8	ped1709
9	ped12320

Figure 8.12 – Near-significant region for CD identified in Multinet_d50 by across-pedigree RGA using case-control ranking
Numbers next to each node refer to the pedigrees in which they contain variants. Jumps were not permitted therefore all genes in the region are seed nodes.

association in the recent IBD GWAS meta-analysis mean that this region should not be prioritised for further study.

8.4 Conclusions

Most research into the genetics of CD is concerned with (not necessarily rare) variants that increase the carrier's risk of disease as part of a multifactorial mechanism. In this chapter we have used whole exome sequencing and network analysis to investigate evidence for an alternative form of CD whose genetic architecture is closer to a monogenic or oligogenic disease, such that one or a small number of rare highly-penetrant variants are responsible for most of the disease risk.

For 25 affected pedigrees we filtered the shared variants identified by whole exome sequencing so that we were left with rare non-synonymous variants in genes that do not frequently contain such variants in non-CD controls. One clear conclusion is that there is no single gene, and no small subnetwork in any of the networks tested, in which all or most of the pedigrees carry a shared post-filtering variant. This suggests that if these families do

have a monogenic or oligogenic form of CD then genetic heterogeneity could present a substantial difficulty in identifying the causal variants.

However, our hypothesis-free (that is, candidate-gene-free) network analyses found several subnetworks for which relevant previous functional annotation suggests that further study could be beneficial. The most promising subnetworks include: the quadruplet *CENPE-MAP3K4-NCF1-TRAF4* in PINA_d50 (Figure 8.4) and several optimal subnetworks in PINAmin2_d50 (Figure 8.6 and Figure 8.7) that were identified by BioGranat-IG (under the assumption of a monogenic cause); the significantly clustered regions found in COXPRES30_d50 using within-pedigree RGA (under the assumption of an oligogenic cause; Figure 8.8); and the region of 31 genes found in PINA_d50 by across-pedigree RGA using case-control gene ranking (Figure 8.11), which includes several proteasome subunit genes and is significantly enriched for rare non-synonymous variants in the CD pedigrees. There is some overlap between these regions, which is not surprising given that there is overlap between the different networks used (see chapter 2) and that, despite the different assumptions they make about the possible genetic architecture of CD, the primary goal of all of the network methods used is to identify connected sets of genes having a concentration of post-filtering sequence variants.

Follow-up work for these regions could involve several steps. Sanger sequencing of the identified variants in these pedigrees would be able to confirm that they had been correctly identified by whole exome sequencing (although the fact that each variant used in these analyses has been independently called in at least two related individuals gives confidence that they are correct). Next, sequencing or genotyping of candidate genes in an additional familial CD cohort could be performed in order to replicate these findings. Over 100 families with multiple CD-affected individuals and a further ~100 individuals with a self-reported family history of CD are available to provide DNA for this purpose at KCL. Finally, strong candidate genes can be subjected to laboratory-based expression and functional testing to establish a biological link to the CD phenotype (such as expression in appropriate tissues, altered expression in mutagenised cell lines or a consistent phenotype in gene knockdown or knockout animal models).

Interestingly none of the genes identified by our analyses interact directly (in any of the networks used) with *NOD2*, currently the most explanatory CD gene (Jostins et al. 2012). *NOD2* is thought to be central to CD biology; for example, in the hypothesised autophagy pathway for CD, the protein encoded by the key CD risk gene *ATG16L1* was shown to be recruited to the cell membrane by NOD2 as a bacterial response (Cooney et al. 2010; Travassos et al. 2010). (Note, however, that this relationship is not represented by an

interaction in any of the networks studied here.) Interaction with *NOD2* would therefore be seen as strong supporting evidence of a CD role.

It should be noted that this study has a number of limitations. A major drawback is that network coverage of the genome is still limited. For example, of the 1,602 genes in which a post-filtering variant is found for at least one pedigree, only 842 (52.6%) are represented in the main interaction network considered here, PINA_d50. Therefore we cannot yet definitively conclude that no small set of interacting genes is responsible for CD in all or most of these families, and it may be worth repeating these analyses when more comprehensive interaction networks become available.

We might also have missed true causal variants due to incomplete sequencing coverage (see Table 8.1). This is of particular relevance in this study where a sequence variant needs to be identified independently in multiple related individuals to be included in our analyses. A more sophisticated approach that might partially overcome this problem could be to perform sequence-based identity-by-descent analysis (Su et al. 2012; Zhuang et al. 2012) within pedigrees to find shared chromosomal segments, and to add to the analysis variants found in at least one individual in any of these regions.

Other limitations are common to the usual intersection filtering approach to whole exome sequence data analysis. We assume that if there are rare highly-penetrant variants underlying familial CD in these pedigrees, these variants will be found in coding regions; in practice variants in introns, or within regulatory elements such as promoters, enhancers and silencers, can also cause inherited disease (Cooper et al. 2010). We also assume that the filtering steps that we take to reduce the number of variants under consideration (excluding relatively common variants, synonymous variants, and variants in genes that are tolerant of rare non-synonymous variants) are appropriate. This could also be incorrect and it is of course possible to repeat all of the analyses using different filtering criteria.

We have tested the hypothesis that CD in these families is caused by high-impact perturbations to specific molecular processes, with the consequence that CD can be effectively thought of as a monogenic or oligogenic disease in these cases. This will almost certainly be an oversimplification of the true disease mechanism. For example, several of the pedigrees may also carry common (known or as-yet-unknown) risk variants for CD; indeed, we know that this is the case for several of the known *NOD2* risk variants (see section 8.2.2). If one or a small number of variants is chiefly responsible for CD in these pedigrees, it is likely that these common risk variants could also play a role as modifying mutations. As described in section 8.2.1, a small number of the CD cases considered were comorbid for other inflammatory or immune-mediated phenotypes, and a small number of CD-unaffected family members were affected by such phenotypes, including UC. The relevance of these

other phenotypes to the molecular basis of CD in these cases is undetermined; due to their clinical heterogeneity a pedigree-by-pedigree investigation of these cases would be warranted to consider possible disease interactions. Finally it should also be acknowledged that there could be a risk of ascertainment bias. Since the common complex form of CD has a substantial genetic component, relatives of CD-affected individuals are at increased risk of developing CD, and the families recruited for this study could represent the unfortunate few where this increased risk is borne out in multiple individuals.

In summary, while the pedigrees studied in this work show strong patterns of inheritance for CD, there is no single gene that is responsible for causing familial CD. CD is a complex disorder and these familial cases are at best likely to have one or a few primary causal variants that sharply increase disease risk alongside smaller effects from many other genetic and non-genetic factors. By undertaking network analyses that examine several potential genetic architectures we make an important step towards elucidating these mechanisms. These analyses have proposed several network regions that could represent key functional pathways that are disrupted in familial CD.

9 Concluding Discussion

9.1 Summary of Findings

Genetic heterogeneity presents a considerable challenge to next generation sequencing studies of rare disease, and there is a need for novel analysis tools to address this problem. I showed in chapter 3 that many monogenic diseases can be caused by different genes that are connected in interaction networks. This motivated the development of two bioinformatic tools, each taking very different approaches to the problem but both making the broad assumption that for a given disease the ability to successfully identify true disease-causing variants among the many identified by whole exome sequencing can be improved by considering genes that are proximal in an interaction network.

BioGranat-IG does this by taking an approach that is very similar to intersection filtering, but does not use individual genes as discrete genomic units in which to look for intersection. Instead, by solving the minimal connected set cover (MCSC) problem BioGranat-IG identifies small connected subnetworks of genes that contain post-filtering variants in all or most exomes in a study. In chapter 4 I simulated data for two real diseases and verified that this concept could effectively overcome genetic heterogeneity. Using synthetic data I also confirmed that BioGranat-IG is most effective when locus heterogeneity is modelled by a disease subnetwork that is small, has relatively low network connectivity and harbours disease-causing variants in the majority of exomes in the study.

HetRank represents a more radical departure from the intersection filtering method; sequence variants are ranked instead of filtered against expected pathogenicity criteria with ranks being adjusted to account for possible locus heterogeneity based on network data. At the time HetRank was developed, experience with real whole exome sequencing data suggested that the simulated data used in chapter 4 to test BioGranat-IG might have been relatively optimistic about the degree of genetic heterogeneity underlying many rare monogenic diseases. In chapter 5 I therefore strove to construct a more realistic set of test data, simulating 1,000 exome sequencing studies of 20 exomes each based on real sequence data and allowing a range of genetic heterogeneity to be modelled. I showed that HetRank was able to prioritise disease-causing genes, achieving a substantial improvement over intersection filtering and BioGranat-IG at higher levels of heterogeneity.

How successfully do network-based methods such as BioGranat-IG and HetRank overcome the limitations of intersection filtering that are imposed by the assumptions discussed in the introduction (section 1.3.4)? These were:

- ***That the disease-causing variants occur in exons.*** This is of course a fundamental constraint of any study limited to whole exome sequencing data and thus applies to both methods. However, as we gain increased understanding of the diverse functional roles of non-coding genomic regions, intersection filtering itself may in future be applicable to whole genome sequencing data. If so, there will undoubtedly also be a role for methods such as these, making use of networks that systematically describe interactions between genomic regions of discrete function (e.g. Gerstein et al. 2012).

- ***That the exonic variants are identified by whole exome sequencing***. For performance-testing of both BioGranat-IG and HetRank, simulated data was designed to model a range of scenarios in which the true disease-causing variants were missing from genes in the interaction network (parameter *p* in chapter 4; "uncaptured heterogeneity" *u* in chapter 5). This could equally well represent imperfect sequencing coverage as imperfect network coverage, with results suggesting that network-based methods may be better equipped than intersection filtering to deal with a small amount of missing data in the presence of locus heterogeneity. Naturally all methods are better powered to prioritise the true causal variants if they are present in whole exome sequencing output.

- ***That the filters applied are appropriate (e.g. for rare, non-synonymous variants or expected mode of inheritance)***. BioGranat-IG, as a generalisation of intersection filtering, is still subject to this assumption. HetRank, however, uses variant-ranking to dispense with the need for filtering, which removes the risk of excluding true causal variants. (But note that for this method we still must assume that our ranking criteria are appropriately chosen and weighted.)

- ***That a single gene is responsible for all or most cases of the disease***. Network-based methods explicitly challenge this assumption. They suppose that disease phenotypes result from the disruption of a functional pathway of interacting gene products and that the information encoded by interaction networks can help to identify and understand these pathways (e.g. Oti and Brunner 2007; Barabasi et al. 2011). This idea is central to both BioGranat-IG and HetRank.

The subsequent chapters have been concerned with the application of BioGranat-IG, HetRank and other network-based methods to real diseases where an intersection filtering approach had previously been unable to identify strong candidate disease genes. All of the methods investigated allowed an exploration of possible mechanisms of locus heterogeneity. In chapter 7 BioGranat-IG and HetRank were employed to study Adams-Oliver syndrome (AOS), an archetypal rare monogenic disease with a heterogeneous genetic basis. Both methods, which are hypothesis-free in the sense that they are exome-wide, highlighted genes with plausible functional links to the disease phenotype and there was some overlap in results. However, arguably the most promising approach was to examine the network neighbourhoods of known disease genes. In chapter 8 familial cases of a complex disorder, Crohn's disease (CD), were studied in order to test the hypothesis that these might be rare, effectively mono- or oligogenic forms of the disease. BioGranat-IG analysis examined the evidence for a monogenic disease mechanism under locus heterogeneity, while two different applications of Region Growing Analysis (RGA; an existing tool of which an improved version was implemented in chapter 6) went beyond this to explore whether a multifactorial genetic architecture, characterised by oligogenicity and genetic heterogeneity, can be elucidated using interaction networks.

Despite suggesting intriguing avenues for further study in each disease, these studies fell short of identifying a "smoking gun" in either case. One reason for this could be the level of genetic heterogeneity underlying each disease, which could make the true causal variants indistinguishable from background variation even when some of these variants occur in interacting genes. If this is the case then larger sample sizes would be expected to provide better power to identify these disease-causing mechanisms using network-based methods (for example, this was demonstrated using simulated data for BioGranat-IG in chapter 4).

Due to the large number of variants identified by whole exome sequencing, one objective of the methods developed here was to prioritise a small number for further study. The range of networks used and analyses performed, however, resulted in a considerable number of subnetworks demanding further attention, albeit ones whose interactions should relatively quickly allow their plausibility as disease mechanisms to be assessed by researchers specialising in these diseases. (But note that for the preliminary assessments I have presented in chapters 7 and 8, existing functional annotation was used as one method of evaluating candidate disease subnetworks; it should be noted that the very use of interaction networks is to facilitate *de novo* pathway discovery (Lehne and Schlitt 2012).)

A desirable property of a gene prioritisation method would be to generate a p-value indicating how unlikely the observed results would be under some null hypothesis (as, for

example, produced by an association test). This would be problematic for the methods described here due to the difficulties in specifying an appropriate null distribution in light of all the variability involved, which includes network structure, sequencing coverage, the propensity of different genes to tolerate functional variation and the necessarily unknown degree of genetic heterogeneity of the disease under study. In addition, BioGranat-IG and HetRank are designed to be used for rare disease studies with relatively small sample size, making them particularly susceptible to confounding statistical noise. The methods should therefore be thought of, like intersection filtering itself, as explorative tools to generate hypotheses rather than statistical methods that test them.

Finally, it is important to be aware that the networks used in this thesis provide a static (and incomplete) representation of an interactome viewed at the whole-organism level. In reality genes are expressed at different levels in different tissues, and these change over time (for example, during embryonic development). To fully understand the functional pathways underlying a disease would require knowledge of the molecular interactions that occur in disease-relevant tissues at appropriate time points. However, until the aetiology of a disease is understood, the most relevant tissues and time points may not be known. Therefore the networks that we use provide a proxy, albeit most likely a poor one, for the "real" interaction network in the body.

9.2 Future Work

9.2.1 BioGranat-IG

In chapter 4 it was suggested (as felt at the time of development) that further work to improve the performance of the heuristic minimum- and multi-minimum distance searches would be beneficial (improved performance here equating to an increased probability that all true optimal MCSC solutions are identified by the algorithms). However, the findings of chapter 7, where BioGranat-IG was applied to real disease data, suggested that the results of the heuristic searches (which by design continue extending candidate subnetworks until they harbour a post-filtering variant for all exomes) may be less informative than the results of the exact triplet and quadruplet searches. I would therefore argue that further development efforts for BioGranat-IG should focus on improving the efficiency of the exact searches. For example, one possible starting point to improve running time might be to work through network genes in descending order of the number of exomes in which they contain variants, examining the triplets and quadruplets they occur in, and stopping this process when it becomes impossible to find a better subnetwork than one already found (a similar short-cut

is already taken by the heuristic searches). Inefficiencies in memory usage should also be addressed, for example due to the accumulation of triplets and quadruplets that are not optimal or within the limits allowed by the user-specified flexibility parameters.

One advantage of improved efficiency would be a better ability to perform searches quickly for highly-connected networks. For example, we saw in chapter 5 (for the comparison against HetRank) that the current implementation was unable to provide triplet search results from 1,000 simulated studies in the full (hub-retained) PINAmin2 network. It might also make it more feasible to estimate the significance of the observed results by permutation testing. This would itself require development of a more sophisticated significance test; the relatively crude test implemented in chapter 4 makes the oversimplifying assumption that network genes are equally likely to contain post-filtering sequence variants in case exomes. One option would be to estimate each gene's propensity to contain variants using control exomes; another is to use the original data in a permuted network (but note this would fix the number of genes containing variants in multiple exomes).

Whether or not improvements to efficiency or algorithm accuracy are made, BioGranat-IG in its present form assumes that solving the discrete unweighted MCSC problem can usefully highlight possible disease-causing pathways based on the presence or absence of variants, and I have tested this concept thoroughly in this thesis. But it would be interesting to consider whether we can move beyond the intrinsic limitations of this model. For example, a formulation of the problem that builds weights into the measure of "fitness" of a candidate subnetwork (currently based on size and number of exomes covered only) might be more effective. Node weights might represent the ability of each gene to tolerate functional variation (thus avoiding the need to exclude variants in highly polymorphic genes as is done in chapters 7 and 8) or a pathogenicity score from a variant effect prediction tool (so that for example the variant scores used in KGGSeq-prioritisation, introduced in chapter 6, are incorporated into the identification of optimal subnetworks and not just the retrospective evaluation of them). Edge weights might reflect interaction confidence or the degree to which interactions are disrupted by the variants in case exomes (assuming the availability of a suitable genome-wide reference database (see e.g. Zhong et al. 2009; Wang et al. 2012)).

9.2.2 HetRank

Of the two methods developed in this thesis, HetRank requires the most additional development to improve its effectiveness at prioritising genes for further study in the

presence of genetic heterogeneity. Application of HetRank to the AOS exomes in chapter 7 revealed a number of features that should be addressed in future iterations of the tool.

A key problem is to redesign the network-based rank adjustment to have a more neutral effect on final gene ranks. At present genes are conferred a considerable ranking advantage simply by being present in the network; this conflicts with one of the original design principles of HetRank: that network-informed ranking would replace subnetwork identification so as not to prevent non-network genes from being prioritised. It is also the case that most of the genes that HetRank prioritised had had their ranks adjusted by better-ranked neighbours in the majority of exomes. While this might be expected due to the presumably small number of actual causal variants and large network neighbourhoods, many of the genes appeared to be highly-ranked on the basis of indirect evidence from network neighbours alone. To be able to demonstrate a causal relationship it is necessary that at least some of the case exomes in a study have direct evidence in the form of a causal variant, and a reformulation of the adjustment mechanism should also try to address this imbalance.

The user-specified weight parameters that HetRank uses to combine variant ranks from various criteria into a final gene rank for each exome should also be considered further. As with any exploratory tool it is important that these parameters can be adjusted to allow different hypotheses to be considered. For example, when the tool was applied to AOS data in chapter 7 we saw that a carefully-reasoned change to the weights – reflecting the belief that causal AOS variants should not be present in other exomes in our in-house database – resulted in genes with more-feasible pathogenic variants being prioritised (see section 7.3.9.2). However, a more systematic approach to calibrating these parameters is desirable, allowing a user to easily quantify their beliefs about the genetic architecture of the disease being studied. This is particularly important when new ranking criteria are introduced, and while various considerations were discussed in chapter 5, further application of HetRank to real datasets should help to establish this approach.

Ideally, HetRank could suggest an initial set of weights automatically based on the input data. There would be several factors to consider. Since a criterion's ability to discriminate different variants is important, weights might account for the information contained in the various criteria (by estimating a measure such as the information entropy of each ranking factor from the input data (Shannon 1948)). However, it would also be necessary to account for correlation between different ranking criteria, such as different measures of allele frequency. A more sophisticated approach again could see the user choose from a list of common disease models (e.g. "rare autosomal dominant") and assign their ranking criteria to built in "types" (e.g. "allele frequency", "zygosity", "pathogenicity").

Again, extensive testing with real datasets would be required to establish the best way to use this information.

Finally, it would be beneficial to develop and test a more formal method of interpreting the ranked gene list output by HetRank. In chapter 7 several approaches were used to focus on individual genes, including counting the number of exomes in which genes are highly ranked, performing RGA and testing for enrichment of existing functional annotation. Ideally these and other methods should be compared and evaluated to produce a recommended HetRank protocol. One possibility might be to incorporate network permutations that allow gene ranks to be compared against those expected under a null hypothesis of non-informative interactions. As with BioGranat-IG, however, a more efficient implementation of HetRank would be required to make this feasible.

9.2.3 Interaction Networks

One of the biggest limitations of all the network-based methods described in this thesis is the incomplete nature of currently available interaction networks. Performance testing in chapters 4 and 5 clearly demonstrated that BioGranat-IG and HetRank both benefit from better network coverage, and we saw in chapter 7 that the ability of both tools to find known AOS genes was compromised by the absence of interactions reflecting the underlying disease mechanisms in several networks. This problem is one that should be gradually resolved by continuing efforts to generate reference interactomes using high-throughput methodologies (see e.g. CCSB 2014).

Meanwhile there are several measures that can be taken to make more effective use of current interaction networks. Firstly, one problem encountered in chapters 7 and 8 was the interpretation of subnetworks identified in the COXPRES30 and COXPRES30_d50 co-expression networks, due to the fact that interactions represent correlated gene expression rather than direct functional relationships. To make these analyses more useful it would be necessary to investigate methods to systematically interpret such subnetworks. The developers of COXPRESdb suggest that consideration of tissue-specific expression or existing functional annotation (such as Gene Ontology or KEGG annotation) could be usefully employed for this purpose (Obayashi et al. 2008). However, the automation of processes such as these would make a challenging data integration project.

Where BioGranat-IG and RGA have been utilised in this thesis, highly-connected hub genes have been removed from the interaction networks used because they tend to be overrepresented in results. However, a pragmatic approach was taken by using an arbitrary degree threshold to select hubs for removal. Identification of a more sophisticated strategy for hub removal would represent a substantial undertaking in itself. One way this could be

done would be to use simulated exome data (as generated in chapter 5, for example) to systematically test the ability of BioGranat-IG to correctly identify "spiked" disease subnetworks over a range of hub-removal thresholds in multiple networks. Optimal hub-removal thresholds would then need to be characterised in terms of network properties in order to make this approach applicable to alternative interaction networks that may become available in future.

9.2.4 Adams-Oliver Syndrome and Familial Crohn's Disease

It would of course be beneficial to re-analyse AOS and familial CD whole exome sequencing data if the improvements to the methods and interaction networks described above are implemented. Meanwhile the putative disease-causing genes highlighted in this thesis can be evaluated. This could lead to the analyses being repeated using alternative parameters following feedback from geneticists with expertise in the respective diseases, or to further laboratory-based experiments to validate or reject candidate genes, as discussed more fully in chapters 7 and 8.

9.3 Conclusions

It was already known that because they encode relationships between genes and gene products, interaction networks can be used to help address a range of problems in genetic data analysis. In this thesis we have shown that one viable application is to help overcome the problem of locus heterogeneity in whole exome sequencing studies of rare monogenic disease, a problem that limits simpler approaches such as intersection filtering.

We have seen that network-based methods can suggest novel disease mechanisms for monogenic diseases and can be used to study suspected mono- or oligogenic forms of complex diseases, potentially leading to an improved understanding of the disease biology which could also be relevant to the common multifactorial form of the disease.

Developments in several areas are needed to make these methods more dependable and results more tractable. These include wider and more precise coverage of relevant interactomes by interaction networks and more comprehensive sources of functional annotation data. Additionally, improvements to the design of the tools presented in this chapter should aim to make them more efficient and more accurate. As next generation sequencing leads to the identification of more causal genes for genetically-heterogeneous monogenic diseases, publicly-deposited sequence data will provide valuable benchmark tests for this task.

It is clear that network-based methods cannot substitute for the knowledge and expertise of skilled geneticists in the identification of disease-causing genes. However, when used appropriately they have the potential to direct researchers towards relevant genes and functional pathways, thus hastening improvements in our understanding of disease biology and in the diagnostic and treatment applications that should follow.

References

1000 Genomes Project Consortium (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Abeysinghe, S. S., N. Chuzhanova and D. N. Cooper (2006). "Gross deletions and translocations in human genetic disease." Genome Dyn **1**: 17-34.

Abou Jamra, R., O. Philippe, A. Raas-Rothschild, S. H. Eck, E. Graf, R. Buchert, G. Borck, A. Ekici, F. F. Brockschmidt, M. M. Nothen, A. Munnich, T. M. Strom, A. Reis and L. Colleaux (2011). "Adaptor protein complex 4 deficiency causes severe autosomal-recessive intellectual disability, progressive spastic paraplegia, shy character, and short stature." Am J Hum Genet **88**(6): 788-795.

Adams, F. H. and C. P. Oliver (1945). "Hereditary Deformities in Man: Due to Arrested Development." Journal of Heredity **36**(1): 3-7.

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev (2010). "A method and server for predicting damaging missense mutations." Nat Methods **7**(4): 248-249.

Aerts, S., D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet and Y. Moreau (2006). "Gene prioritization through genomic data fusion." Nat Biotechnol **24**(5): 537-544.

Ajay, S. S., S. C. Parker, H. O. Abaan, K. V. Fajardo and E. H. Margulies (2011). "Accurate and comprehensive sequencing of personal genomes." Genome Res **21**(9): 1498-1505.

Albert, R. (2005). "Scale-free networks in cell biology." J Cell Sci **118**(Pt 21): 4947-4957.

Alcaraz, N., H. Kücük, J. Weile, A. Wipat and J. Baumbach (2011). "KeyPathwayMiner: Detecting Case-Specific Biological Pathways Using Expression Data." Internet Mathematics **7**(4): 299-313.

Alcaraz, N., T. Friedrich, T. Kotzing, A. Krohmer, J. Muller, J. Pauling and J. Baumbach (2012). "Efficient key pathway mining: combining networks and OMICS data." Integr Biol (Camb) **4**(7): 756-764.

Alon, U. (2007). "Network motifs: theory and experimental approaches." Nat Rev Genet **8**(6): 450-461.

Altelaar, A. F., J. Munoz and A. J. Heck (2013). "Next-generation proteomics: towards an integrative view of proteome dynamics." Nat Rev Genet **14**(1): 35-48.

Amberger, J., C. A. Bocchini, A. F. Scott and A. Hamosh (2009). "McKusick's Online Mendelian Inheritance in Man (OMIM)." Nucleic Acids Res **37**(Database issue): D793-796.

Antonarakis, S. E. and J. S. Beckmann (2006). "Mendelian disorders deserve more attention." Nat Rev Genet **7**(4): 277-282.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and

G. Sherlock (2000). "Gene Ontology: tool for the unification of biology." Nat Genet **25**(1): 25-29.

Babu, M. M., N. M. Luscombe, L. Aravind, M. Gerstein and S. A. Teichmann (2004). "Structure and evolution of transcriptional regulatory networks." Curr Opin Struct Biol **14**(3): 283-291.

Backes, C., A. Rurainski, G. W. Klau, O. Muller, D. Stockel, A. Gerasch, J. Kuntzer, D. Maisel, N. Ludwig, M. Hein, A. Keller, H. Burtscher, M. Kaufmann, E. Meese and H. P. Lenhof (2012). "An integer linear programming approach for finding deregulated subgraphs in regulatory networks." Nucleic Acids Res **40**(6): e43.

Bailey, J. A., A. M. Yavor, H. F. Massa, B. J. Trask and E. E. Eichler (2001). "Segmental duplications: organization and impact within the current human genome project assembly." Genome Res **11**(6): 1005-1017.

Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li and E. E. Eichler (2002). "Recent segmental duplications in the human genome." Science **297**(5583): 1003-1007.

Bamshad, M. J., S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson and J. Shendure (2011). "Exome sequencing as a tool for Mendelian disease gene discovery." Nat Rev Genet **12**(11): 745-755.

Barabasi, A. L., N. Gulbahce and J. Loscalzo (2011). "Network medicine: a network-based approach to human disease." Nat Rev Genet **12**(1): 56-68.

Baranzini, S. E., N. W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, W. Wu, B. M. Uitdehaag, L. Kappos, C. H. Polman, P. M. Matthews, S. L. Hauser, R. A. Gibson, J. R. Oksenberg and M. R. Barnes (2009). "Pathway and network-based analysis of genome-wide association studies in multiple sclerosis." Hum Mol Genet **18**(11): 2078-2090.

Barrenas, F., S. Chavali, P. Holme, R. Mobini and M. Benson (2009). "Network properties of complex human disease genes identified through genome-wide association studies." Plos One **4**(11): e8090.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B (Methodological) **57**(1): 289-300.

Blekhman, R., O. Man, L. Herrmann, A. R. Boyko, A. Indap, C. Kosiol, C. D. Bustamante, K. M. Teshima and M. Przeworski (2008). "Natural selection on genes that underlie human disease susceptibility." Curr Biol **18**(12): 883-889.

Boycott, K. M., M. R. Vanstone, D. E. Bulman and A. E. MacKenzie (2013). "Rare-disease genetics in the era of next-generation sequencing: discovery to translation." Nat Rev Genet **14**(10): 681-691.

Brand, S. (2009). "Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease." Gut **58**(8): 1152-1167.

Brant, S. R. (2011). "Update on the heritability of inflammatory bowel disease: the importance of twin studies." Inflamm Bowel Dis **17**(1): 1-5.

Brin, S. and L. Page (1998). "The anatomy of a large-scale hypertextual Web search engine." Computer Networks and ISDN Systems **30**(1–7): 107-117.

Brunham, L. R. and M. R. Hayden (2013). "Hunting human disease genes: lessons from the past, challenges for the future." Hum Genet **132**(6): 603-617.

Canessa, C. M., L. Schild, G. Buell, B. Thorens, I. Gautschi, J. D. Horisberger and B. C. Rossier (1994). "Amiloride-sensitive epithelial Na+ channel is made of three homologous subunits." Nature **367**(6462): 463-467.

Carter, H., C. Douville, P. D. Stenson, D. N. Cooper and R. Karchin (2013). "Identifying Mendelian disease genes with the variant effect scoring tool." Bmc Genomics **14 Suppl 3**: S3.

CCSB (2014). "Human Interactome Project." Centre for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston MA. Retrieved 20/11/2014, from http://interactome.dfci.harvard.edu/H_sapiens/.

Cerdeira, J. O. and L. S. Pinto (2005). "Requiring connectivity in the set covering problem." Journal of Combinatorial Optimization **9**(1): 35-47.

Chatr-Aryamontri, A., B. J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, T. Reguly, A. Breitkreutz, A. Sellam, D. Chen, C. Chang, J. Rust, M. Livstone, R. Oughtred, K. Dolinski and M. Tyers (2013). "The BioGRID interaction database: 2013 update." Nucleic Acids Res **41**(Database issue): D816-823.

Chen, W. V. and T. Maniatis (2013). "Clustered protocadherins." Development **140**(16): 3297-3302.

Chowdhury, S. A., R. K. Nibbe, M. R. Chance and M. Koyuturk (2011). "Subnetwork state functions define dysregulated subnetworks in cancer." J Comput Biol **18**(3): 263-281.

Chuang, H. Y., E. Lee, Y. T. Liu, D. Lee and T. Ideker (2007). "Network-based classification of breast cancer metastasis." Mol Syst Biol **3**: 140.

Clark, M. J., R. Chen, H. Y. Lam, K. J. Karczewski, G. Euskirchen, A. J. Butte and M. Snyder (2011). "Performance comparison of exome DNA sequencing technologies." Nat Biotechnol **29**(10): 908-914.

Cohen, I., E. Silberstein, Y. Perez, D. Landau, K. Elbedour, Y. Langer, R. Kadir, M. Volodarsky, S. Sivan, G. Narkis and O. S. Birk (2014). "Autosomal recessive Adams-Oliver syndrome caused by homozygous mutation in EOGT, encoding an EGF domain-specific O-GlcNAc transferase." Eur J Hum Genet **22**(3): 374-378.

Cohen, P. (2014). "Immune diseases caused by mutations in kinases and components of the ubiquitin system." Nat Immunol **15**(6): 521-529.

Cooney, R., J. Baker, O. Brain, B. Danis, T. Pichulik, P. Allan, D. J. Ferguson, B. J. Campbell, D. Jewell and A. Simmons (2010). "NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation." Nat Med **16**(1): 90-97.

Cooper, D. N., J. M. Chen, E. V. Ball, K. Howells, M. Mort, A. D. Phillips, N. Chuzhanova, M. Krawczak, H. Kehrer-Sawatzki and P. D. Stenson (2010). "Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics." Hum Mutat **31**(6): 631-655.

Cooper, D. N., M. Krawczak, C. Polychronakos, C. Tyler-Smith and H. Kehrer-Sawatzki (2013). "Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease." Hum Genet **132**(10): 1077-1130.

Cooper, G. M., E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou and A. Sidow (2005). "Distribution and intensity of constraint in mammalian genomic sequence." Genome Res **15**(7): 901-913.

Cooper, G. M. and J. Shendure (2011). "Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data." Nat Rev Genet **12**(9): 628-640.

Cormen, T. H. (2001). Introduction to algorithms. Cambridge, Mass., MIT Press.

Corominas, R., X. Yang, G. N. Lin, S. Kang, Y. Shen, L. Ghamsari, M. Broly, M. Rodriguez, S. Tam, S. A. Trigg, C. Fan, S. Yi, M. Tasan, I. Lemmens, X. Kuang, N. Zhao, D. Malhotra, J. J. Michaelson, V. Vacic, M. A. Calderwood, F. P. Roth, J. Tavernier, S. Horvath, K. Salehi-Ashtiani, D. Korkin, J. Sebat, D. E. Hill, T. Hao, M. Vidal and L. M. Iakoucheva (2014). "Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism." Nat Commun **5**: 3650.

Costanzo, M., A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z. Y. Lin, W. Liang, M. Marback, J. Paw, B. J. San Luis, E. Shuteriqi, A. H. Tong, N. van Dyk, et al. (2010). "The genetic landscape of a cell." Science **327**(5964): 425-431.

Cowley, M. J., M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, A. V. Biankin, S. Hautaniemi and J. Wu (2012). "PINA v2.0: mining interactome modules." Nucleic Acids Res **40**(Database issue): D862-865.

Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-563.

Cruchaga, C., C. M. Karch, S. C. Jin, B. A. Benitez, Y. Cai, R. Guerreiro, O. Harari, J. Norton, J. Budde, S. Bertelsen, A. T. Jeng, B. Cooper, T. Skorupa, D. Carrell, D. Levitch, S. Hsu, J. Choi, M. Ryten, U. K. B. E. Consortium, C. Sassi, J. Bras, J. R. Gibbs, D. G. Hernandez, M. K. Lupton, J. Powell, P. Forabosco, P. G. Ridge, C. D. Corcoran, J. T. Tschanz, M. C. Norton, et al. (2014). "Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer/'s disease." Nature **505**(7484): 550-554.

Csardi, G. and T. Nepusz (2006). "The igraph software package for complex network research." InterJournal, Complex Systems **1695**(5).

Cullinane, A. R., T. Vilboux, K. O'Brien, J. A. Curry, D. M. Maynard, H. Carlson-Donohoe, C. Ciccone, T. C. Markello, M. Gunay-Aygun, M. Huizing and W. A. Gahl (2011). "Homozygosity mapping and whole-exome sequencing to detect SLC45A2 and G6PC3 mutations in a single patient with oculocutaneous albinism and neutropenia." J Invest Dermatol **131**(10): 2017-2025.

Dand, N., F. Sprengel, V. Ahlers and T. Schlitt (2013). "BioGranat-IG: a network analysis tool to suggest mechanisms of genetic heterogeneity from exome-sequencing data." Bioinformatics **29**(6): 733-741.

Dao, P., R. Colak, R. Salari, F. Moser, E. Davicioni, A. Schonhuth and M. Ester (2010). "Inferring cancer subnetwork markers using density-constrained biclustering." Bioinformatics **26**(18): i625-631.

den Dunnen, J. T. and S. E. Antonarakis (2000). "Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion." Hum Mutat **15**(1): 7-12.

Depienne, C., D. Bouteiller, A. Meneret, S. Billot, S. Groppa, S. Klebe, F. Charbonnier-Beaupel, J. C. Corvol, J. P. Saraiva, N. Brueggemann, K. Bhatia, M. Cincotta, V. Brochard, C. Flamand-Roze, W. Carpentier, S. Meunier, Y. Marie, M. Gaussen, G. Stevanin, R. Wehrle, M. Vidailhet, C. Klein, I. Dusart, A. Brice and E. Roze (2012). "RAD51 haploinsufficiency causes congenital mirror movements in humans." Am J Hum Genet **90**(2): 301-307.

Dipple, K. M. and E. R. McCabe (2000). "Modifier genes convert "simple" Mendelian disorders to complex traits." Mol Genet Metab **71**(1-2): 43-50.

Dittrich, M. T., G. W. Klau, A. Rosenwald, T. Dandekar and T. Muller (2008). "Identifying functional modules in protein-protein interaction networks: an integrated exact approach." Bioinformatics **24**(13): i223-231.

Djebbari, A. and J. Quackenbush (2008). "Seeded Bayesian Networks: constructing genetic networks from microarray data." BMC Syst Biol **2**: 57.

Drumm, M. L., M. W. Konstan, M. D. Schluchter, A. Handler, R. Pace, F. Zou, M. Zariwala, D. Fargo, A. Xu, J. M. Dunn, R. J. Darrah, R. Dorfman, A. J. Sandford, M. Corey, J. Zielenski, P. Durie, K. Goddard, J. R. Yankaskas, F. A. Wright and M. R. Knowles (2005). "Genetic modifiers of lung disease in cystic fibrosis." N Engl J Med **353**(14): 1443-1453.

Duarte, N. C., S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas and B. O. Palsson (2007). "Global reconstruction of the human metabolic network based on genomic and bibliomic data." Proc Natl Acad Sci U S A **104**(6): 1777-1782.

Duerr, R. H., K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, T. Dassopoulos, A. Bitton, H. Yang, S. Targan, L. W. Datta, E. O. Kistner, L. P. Schumm, A. T. Lee, P. K. Gregersen, M. M. Barmada, J. I. Rotter, D. L. Nicolae and J. H. Cho (2006). "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene." Science **314**(5804): 1461-1463.

Eddy, S. R. (2013). "The ENCODE project: missteps overshadowing a success." Curr Biol **23**(7): R259-261.

Elbassioni, K., S. Jelić and D. Matijević (2012). "The relation of Connected Set Cover and Group Steiner Tree." Theoretical Computer Science **438**(0): 96-101.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Erlich, Y., S. Edvardson, E. Hodges, S. Zenvirt, P. Thekkat, A. Shaag, T. Dor, G. J. Hannon and O. Elpeleg (2011). "Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis." Genome Res **21**(5): 658-664.

Feldman, I., A. Rzhetsky and D. Vitkup (2008). "Network properties of genes harboring inherited disease mutations." Proc Natl Acad Sci U S A **105**(11): 4323-4328.

Fernández-Suárez, X. M., D. J. Rigden and M. Y. Galperin (2014). "The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection." Nucleic Acids Res **42**(Database issue): D1-6.

Fields, S. and O. Song (1989). "A novel genetic system to detect protein-protein interactions." Nature **340**(6230): 245-246.

Firsov, D., L. Schild, I. Gautschi, A. M. Merillat, E. Schneeberger and B. C. Rossier (1996). "Cell surface expression of the epithelial Na channel and a mutant causing Liddle syndrome: a quantitative approach." Proc Natl Acad Sci U S A **93**(26): 15370-15375.

Fisher, R. A. (1932). Statistical methods for research workers. Edinburgh, London, Oliver and Boyd.

Flicek, P. and E. Birney (2009). "Sense from sequence reads: methods for alignment and assembly." Nat Methods **6**(11 Suppl): S6-S12.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kahari, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, et al. (2014). "Ensembl 2014." Nucleic Acids Res **42**(Database issue): D749-755.

Fortunato, S. (2010). "Community detection in graphs." Physics Reports-Review Section of Physics Letters **486**(3-5): 75-174.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. J. Jensen (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." Nucleic Acids Res **41**(Database issue): D808-815.

Franke, A., D. P. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, C. A. Anderson, J. C. Bis, S. Bumpstead, D. Ellinghaus, E. M. Festen, M. Georges, T. Green, T. Haritunians, L. Jostins, A. Latiano, C. G. Mathew, G. W. Montgomery, N. J. Prescott, S. Raychaudhuri, J. I. Rotter, P. Schumm, Y. Sharma, L. A. Simms, K. D. Taylor, D. Whiteman, et al. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." Nat Genet **42**(12): 1118-1125.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe and M. W. Feldman (2002). "Evolutionary rate in the protein interaction network." Science **296**(5568): 750-752.

Frodsham, A. J. and A. V. Hill (2004). "Genetics of infectious diseases." Hum Mol Genet **13 Spec No 2**: R187-194.

Fromer, M., A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, N. Carrera, I. Humphreys, J. S. Johnson, P. Roussos, D. D. Barker, E. Banks, V. Milanova, S. G. Grant, E. Hannon, S. A. Rose, K. Chambert, M. Mahajan, E. M. Scolnick, J. L. Moran, G. Kirov, A. Palotie, S. A. McCarroll, P. Holmans, P. Sklar, M. J. Owen, et al. (2014). "De novo mutations in schizophrenia implicate synaptic networks." Nature **advance online publication**.

Frousios, K., C. S. Iliopoulos, T. Schlitt and M. A. Simpson (2013). "Predicting the functional consequences of non-synonymous DNA sequence variants--evaluation of bioinformatics tools and development of a consensus strategy." Genomics **102**(4): 223-228.

Fuentes Fajardo, K. V., D. Adams, C. E. Mason, M. Sincan, C. Tifft, C. Toro, C. F. Boerkoel, W. Gahl and T. Markello (2012). "Detecting false-positive signals in exome sequencing." Hum Mutat **33**(4): 609-613.

Gamage, N., A. Barnett, N. Hempel, R. G. Duggleby, K. F. Windmill, J. L. Martin and M. E. McManus (2006). "Human sulfotransferases and their role in chemical metabolism." Toxicol Sci **90**(1): 5-22.

Gandhi, T. K., J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nandakumar, B. Shen, N. Deshpande, R. Nayak, M. Sarker, J. D. Boeke, G. Parmigiani, J. Schultz, J. S. Bader and A. Pandey (2006). "Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets." Nat Genet **38**(3): 285-293.

Garcia-Alonso, L., R. Alonso, E. Vidal, A. Amadoz, A. de Maria, P. Minguez, I. Medina and J. Dopazo (2012). "Discovering the hidden sub-network component in a ranked list

of genes or proteins derived from genomic experiments." Nucleic Acids Res **40**(20): e158.

Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, et al. (2012). "Architecture of the human regulatory network derived from ENCODE data." Nature **489**(7414): 91-100.

Gibson, G. (2011). "Rare and common variants: twenty arguments." Nat Rev Genet **13**(2): 135-145.

Gilissen, C., A. Hoischen, H. G. Brunner and J. A. Veltman (2011). "Unlocking Mendelian disease using exome sequencing." Genome Biol **12**(9): 228.

Gillis, J. and P. Pavlidis (2011). "The impact of multifunctional genes on "guilt by association" analysis." Plos One **6**(2): e17258.

Goh, K. I., M. E. Cusick, D. Valle, B. Childs, M. Vidal and A. L. Barabasi (2007). "The human disease network." Proc Natl Acad Sci U S A **104**(21): 8685-8690.

Goh, L., G. B. Chen, I. Cutcutache, B. Low, B. T. Teh, S. Rozen and P. Tan (2011). "Assessing matched normal and tumor pairs in next-generation sequencing studies." Plos One **6**(3): e17810.

Gonzalez-Perez, A. and N. Lopez-Bigas (2011). "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel." Am J Hum Genet **88**(4): 440-449.

Gray, K. A., L. C. Daugherty, S. M. Gordon, R. L. Seal, M. W. Wright and E. A. Bruford (2013). "Genenames.org: the HGNC resources in 2013." Nucleic Acids Res **41**(Database issue): D545-552.

Guo, Y., J. Long, J. He, C. I. Li, Q. Cai, X. O. Shu, W. Zheng and C. Li (2012). "Exome sequencing generates high quality data in non-target regions." Bmc Genomics **13**: 194.

Hahn, M. W. and A. D. Kern (2005). "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks." Mol Biol Evol **22**(4): 803-806.

Halme, L., P. Paavola-Sakki, U. Turunen, M. Lappalainen, M. Farkkila and K. Kontula (2006). "Family and twin studies in inflammatory bowel disease." World J Gastroenterol **12**(23): 3668-3672.

Hampe, J., A. Franke, P. Rosenstiel, A. Till, M. Teuber, K. Huse, M. Albrecht, G. Mayr, F. M. De La Vega, J. Briggs, S. Gunther, N. J. Prescott, C. M. Onnie, R. Hasler, B. Sipos, U. R. Folsch, T. Lengauer, M. Platzer, C. G. Mathew, M. Krawczak and S. Schreiber (2007). "A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1." Nat Genet **39**(2): 207-211.

Han, J. D., N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth and M. Vidal (2004). "Evidence for dynamically organized modularity in the yeast protein-protein interaction network." Nature **430**(6995): 88-93.

Hassed, S. J., G. B. Wiley, S. Wang, J. Y. Lee, S. Li, W. Xu, Z. J. Zhao, J. J. Mulvihill, J. Robertson, J. Warner and P. M. Gaffney (2012). "RBPJ mutations identified in two families affected by Adams-Oliver syndrome." Am J Hum Genet **91**(2): 391-395.

Hatem, A., D. Bozdag, A. E. Toland and U. V. Catalyurek (2013). "Benchmarking short sequence mapping tools." BMC Bioinformatics **14**: 184.

Hawkins, R. D., G. C. Hon and B. Ren (2010). "Next-generation genomics: an integrative approach." Nat Rev Genet **11**(7): 476-486.

He, X. and J. Zhang (2006). "Why do hubs tend to be essential in protein networks?" PLoS Genet **2**(6): e88.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proc Natl Acad Sci U S A **106**(23): 9362-9367.

Hoischen, A., B. W. van Bon, C. Gilissen, P. Arts, B. van Lier, M. Steehouwer, P. de Vries, R. de Reuver, N. Wieskamp, G. Mortier, K. Devriendt, M. Z. Amorim, N. Revencu, A. Kidd, M. Barbosa, A. Turner, J. Smith, C. Oley, A. Henderson, I. M. Hayes, E. M. Thompson, H. G. Brunner, B. B. de Vries and J. A. Veltman (2010). "De novo mutations of SETBP1 cause Schinzel-Giedion syndrome." Nat Genet **42**(6): 483-485.

Holmen, O. L., H. Zhang, Y. Fan, D. H. Hovelson, E. M. Schmidt, W. Zhou, Y. Guo, J. Zhang, A. Langhammer, M. L. Lochen, S. K. Ganesh, L. Vatten, F. Skorpen, H. Dalen, S. Pennathur, J. Chen, C. Platou, E. B. Mathiesen, T. Wilsgaard, I. Njolstad, M. Boehnke, Y. E. Chen, G. R. Abecasis, K. Hveem and C. J. Willer (2014). "Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk." Nat Genet **46**(4): 345-351.

Hood, R. L., M. A. Lines, S. M. Nikkel, J. Schwartzentruber, C. Beaulieu, M. J. Nowaczyk, J. Allanson, C. A. Kim, D. Wieczorek, J. S. Moilanen, D. Lacombe, G. Gillessen-Kaesbach, M. L. Whiteford, C. R. Quaio, I. Gomy, D. R. Bertola, B. Albrecht, K. Platzer, G. McGillivray, R. Zou, D. R. McLeod, A. E. Chudley, B. N. Chodirker, J. Marcadier, J. Majewski, D. E. Bulman, S. M. White and K. M. Boycott (2012). "Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome." Am J Hum Genet **90**(2): 308-313.

Hugot, J. P., M. Chamaillard, H. Zouali, S. Lesage, J. P. Cezard, J. Belaiche, S. Almer, C. Tysk, C. A. O'Morain, M. Gassull, V. Binder, Y. Finkel, A. Cortot, R. Modigliani, P. Laurent-Puig, C. Gower-Rousseau, J. Macry, J. F. Colombel, M. Sahbatou and G. Thomas (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.

Ideker, T., O. Ozier, B. Schwikowski and A. F. Siegel (2002). "Discovering regulatory and signalling circuits in molecular interaction networks." Bioinformatics **18 Suppl 1**: S233-240.

Ideker, T. and N. J. Krogan (2012). "Differential network biology." Mol Syst Biol **8**: 565.

Illumina (2014a). "HiSeq X Ten." Retrieved 09/05/2014, from http://www.illumina.com/systems/hiseq-x-sequencing-system.ilmn.

Illumina (2014b). "Omni Array Family." Retrieved 30/11/2014, from http://applications.illumina.com/applications/genotyping/human-genotyping-arrays/omni-arrays.html.

Isserlin, R., R. A. El-Badrawi and G. D. Bader (2011). "The Biomolecular Interaction Network Database in PSI-MI 2.5." Database **2011**.

Jenssen, T. K., A. Laegreid, J. Komorowski and E. Hovig (2001). "A literature network of human genes for high-throughput analysis of gene expression." Nat Genet **28**(1): 21-28.

Jeong, H., S. P. Mason, A. L. Barabasi and Z. N. Oltvai (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-42.

Jia, P. and Z. Zhao (2014). "VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data." PLoS Comput Biol **10**(2): e1003460.

Johnson, R. C., G. W. Nelson, J. L. Troyer, J. A. Lautenberger, B. D. Kessing, C. A. Winkler and S. J. O'Brien (2010). "Accounting for multiple comparisons in a genome-wide association study (GWAS)." Bmc Genomics **11**: 724.

Johnston, J. J., J. K. Teer, P. F. Cherukuri, N. F. Hansen, S. K. Loftus, K. Chong, J. C. Mullikin and L. G. Biesecker (2010). "Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate." Am J Hum Genet **86**(5): 743-748.

Jones, W. D., D. Dafou, M. McEntagart, W. J. Woollard, F. V. Elmslie, M. Holder-Espinasse, M. Irving, A. K. Saggar, S. Smithson, R. C. Trembath, C. Deshpande and M. A. Simpson (2012). "De novo mutations in MLL cause Wiedemann-Steiner syndrome." Am J Hum Genet **91**(2): 358-364.

Jostins, L., S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J. P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, et al. (2012). "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease." Nature **491**(7422): 119-124.

Kamburov, A., U. Stelzl and R. Herwig (2012). "IntScore: a web tool for confidence scoring of biological interactions." Nucleic Acids Res **40**(Web Server issue): W140-146.

Kamburov, A., U. Stelzl, H. Lehrach and R. Herwig (2013). "The ConsensusPathDB interaction database: 2013 update." Nucleic Acids Res **41**(Database issue): D793-800.

Karolchik, D., G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn and W. J. Kent (2014). "The UCSC Genome Browser database: 2014 update." Nucleic Acids Res **42**(Database issue): D764-770.

Karp, R. (1972). Reducibility among combinatorial problems. Complexity of Computer Computations. R. Miller and J. Thatcher, Plenum Press**:** 85--103.

Kasprzyk, A. (2011). "BioMart: driving a paradigm change in biological data management." Database (Oxford) **2011**: bar049.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Keshava Prasad, T. S., R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J.

Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey (2009). "Human Protein Reference Database—2009 update." Nucleic Acids Res **37**(suppl 1): D767-D772.

Khor, B., A. Gardet and R. J. Xavier (2011). "Genetics and pathogenesis of inflammatory bowel disease." Nature **474**(7351): 307-317.

Khurana, E., Y. Fu, J. Chen and M. Gerstein (2013). "Interpretation of genomic variants using a unified biological network approach." PLoS Comput Biol **9**(3): e1002886.

Kim, P. M., L. J. Lu, Y. Xia and M. B. Gerstein (2006). "Relating three-dimensional structures to protein networks provides evolutionary insights." Science **314**(5807): 1938-1941.

Kim, P. M., J. O. Korbel and M. B. Gerstein (2007). "Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context." Proc Natl Acad Sci U S A **104**(51): 20274-20279.

Kim, W., M. Li, J. Wang and Y. Pan (2011). "Biological network motif detection and evaluation." BMC Syst Biol **5 Suppl 3**: S5.

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper and J. Shendure (2014). "A general framework for estimating the relative pathogenicity of human genetic variants." Nat Genet **46**(3): 310-315.

Kirkpatrick, S., C. D. Gelatt, Jr. and M. P. Vecchi (1983). "Optimization by simulated annealing." Science **220**(4598): 671-680.

Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable and J. Hoh (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-389.

Koboldt, D. C., L. Ding, E. R. Mardis and R. K. Wilson (2010). "Challenges of sequencing human genomes." Brief Bioinform **11**(5): 484-498.

Kondoh, K., K. Sunadome and E. Nishida (2007). "Notch signaling suppresses p38 MAPK activity via induction of MKP-1 in myogenesis." J Biol Chem **282**(5): 3058-3065.

Krueger, F., B. Kreck, A. Franke and S. R. Andrews (2012). "DNA methylome analysis using short bisulfite sequencing data." Nat Methods **9**(2): 145-151.

Ku, C. S., N. Naidoo and Y. Pawitan (2011). "Revisiting Mendelian disorders through exome sequencing." Hum Genet **129**(4): 351-370.

Kuhlenbaumer, G., J. Hullmann and S. Appenzeller (2011). "Novel genomic techniques open new avenues in the analysis of monogenic disorders." Hum Mutat **32**(2): 144-151.

Kumar, P., S. Henikoff and P. C. Ng (2009). "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm." Nat Protoc **4**(7): 1073-1081.

Küster, W., W. Lenz, H. Kaariainen and F. Majewski (1988). "Congenital scalp defects with distal limb anomalies (Adams-Oliver syndrome): report of ten cases and review of the literature." Am J Med Genet **31**(1): 99-115.

Lalonde, E., S. Albrecht, K. C. Ha, K. Jacob, N. Bolduc, C. Polychronakos, P. Dechelotte, J. Majewski and N. Jabado (2010). "Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing." Hum Mutat **31**(8): 918-923.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Landrum, M. J., J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church and D. R. Maglott (2014). "ClinVar: public archive of relationships among sequence variation and human phenotype." Nucleic Acids Res **42**(Database issue): D980-985.

Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." BMC Bioinformatics **9**: 559.

Lee, I., U. M. Blom, P. I. Wang, J. E. Shim and E. M. Marcotte (2011). "Prioritizing candidate disease genes by network-based boosting of genome-wide association data." Genome Research **21**(7): 1109-1121.

Lee, S., M. C. Wu and X. Lin (2012). "Optimal tests for rare variant effects in sequencing association studies." Biostatistics **13**(4): 762-775.

Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.

Lehne, B. and T. Schlitt (2009). "Protein-protein interaction databases: keeping up with growing interactomes." Hum Genomics **3**(3): 291-297.

Lehne, B. (2011). Computational Analyses of Complex Diseases at the Gene and Network Levels. PhD Thesis, King's College London, UK.

Lehne, B. and T. Schlitt (2012). "Breaking free from the chains of pathway annotation: de novo pathway discovery for the analysis of disease processes." Pharmacogenomics **13**(16): 1967-1978.

Lewis, A. C., N. S. Jones, M. A. Porter and C. M. Deane (2010). "The function of communities in protein interaction networks at multiple scales." BMC Syst Biol **4**: 100.

Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." Am J Hum Genet **83**(3): 311-321.

Li, H., J. Ruan and R. Durbin (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res **18**(11): 1851-1858.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, H. and N. Homer (2010). "A survey of sequence alignment algorithms for next-generation sequencing." Brief Bioinform **11**(5): 473-483.

Li, J., Y. Liu, M. Liu and J. D. Han (2013). "Functional dissection of regulatory models using gene expression data of deletion mutants." PLoS Genet **9**(9): e1003757.

Li, M. X., H. S. Gui, J. S. Kwan, S. Y. Bao and P. C. Sham (2012). "A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases." Nucleic Acids Res **40**(7): e53.

Li, X., M. Wu, C. K. Kwoh and S. K. Ng (2010). "Computational approaches for detecting protein complexes from protein interaction networks: a survey." Bmc Genomics **11 Suppl 1**: S3.

Lines, M. A., L. Huang, J. Schwartzentruber, S. L. Douglas, D. C. Lynch, C. Beaulieu, M. L. Guion-Almeida, R. M. Zechi-Ceide, B. Gener, G. Gillessen-Kaesbach, C. Nava, G. Baujat, D. Horn, U. Kini, A. Caliebe, Y. Alanay, G. E. Utine, D. Lev, J. Kohlhase, A. W. Grix, D. R. Lohmann, U. Hehr, D. Bohm, J. Majewski, D. E. Bulman, D. Wieczorek and K. M. Boycott (2012). "Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly." Am J Hum Genet **90**(2): 369-377.

Liu, G., C. H. Yong, H. N. Chua and L. Wong (2011). "Decomposing PPI networks for complex discovery." Proteome Sci **9 Suppl 1**: S15.

Liu, J. Z. and C. A. Anderson (2014). "Genetic studies of Crohn's disease: past, present and future." Best Pract Res Clin Gastroenterol **28**(3): 373-386.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law (2012). "Comparison of next-generation sequencing systems." J Biomed Biotechnol **2012**: 251364.

Liu, X., S. Han, Z. Wang, J. Gelernter and B. Z. Yang (2013). "Variant callers for next-generation sequencing data: a comparison study." Plos One **8**(9): e75619.

Lobo, I. (2008). "Same genetic mutation, different genetic disease phenotype." Nature Education **1**(1): 64.

Loftus, E. V., Jr. (2004). "Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences." Gastroenterology **126**(6): 1504-1517.

Lohmueller, K. E., T. Sparso, Q. Li, E. Andersson, T. Korneliussen, A. Albrechtsen, K. Banasik, N. Grarup, I. Hallgrimsdottir, K. Kiil, T. O. Kilpelainen, N. T. Krarup, T. H. Pers, G. Sanchez, Y. Hu, M. Degiorgio, T. Jorgensen, A. Sandbaek, T. Lauritzen, S. Brunak, K. Kristiansen, Y. Li, T. Hansen, J. Wang, R. Nielsen and O. Pedersen (2013). "Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes." Am J Hum Genet **93**(6): 1072-1086.

Lopes, M. C., C. Joyce, G. R. Ritchie, S. L. John, F. Cunningham, J. Asimit and E. Zeggini (2012). "A combined functional annotation score for non-synonymous variants." Hum Hered **73**(1): 47-51.

Luo, F., Y. Yang, C. F. Chen, R. Chang, J. Zhou and R. H. Scheuermann (2007). "Modular organization of protein interaction networks." Bioinformatics **23**(2): 207-214.

Lupski, J. R. (1998). "Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits." Trends Genet **14**(10): 417-422.

MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler and C. Gunter (2014). "Guidelines for investigating causality of sequence variants in human disease." Nature **508**(7497): 469-476.

Madsen, B. E. and S. R. Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic." PLoS Genet **5**(2): e1000384.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure and D. J. Turner (2010). "Target-enrichment strategies for next-generation sequencing." Nat Methods **7**(2): 111-118.

Mardis, E. R. (2011). "A decade's perspective on DNA sequencing technology." Nature **470**(7333): 198-203.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res **18**(9): 1509-1517.

Mason, O. and M. Verwoerd (2007). "Graph theory and networks in Biology." Systems Biology, IET **1**(2): 89-119.

Mathew, C. G. (2008). "New links to the pathogenesis of Crohn disease provided by genome-wide association scans." Nat Rev Genet **9**(1): 9-14.

Matricon, J., N. Barnich and D. Ardid (2010). "Immunopathogenesis of inflammatory bowel disease." Self Nonself **1**(4): 299-309.

McClellan, J. and M. C. King (2010). "Genetic heterogeneity in human disease." Cell **141**(2): 210-217.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. A. DePristo (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-1303.

Mendig, A., F. Sprengel, T. Schlitt and V. Ahlers (2009). GPU-beschleunigtes 3D-Layout komplexer Netzwerke. Go-3D 2009: Go for Innovations, Stuttgart, Fraunhofer-Verlag.

Mertes, F., A. Elsharawy, S. Sauer, J. M. van Helvoort, P. J. van der Zaag, A. Franke, M. Nilsson, H. Lehrach and A. J. Brookes (2011). "Targeted enrichment of genomic DNA regions for next-generation sequencing." Brief Funct Genomics **10**(6): 374-386.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Milenković, T. and N. Pržulj (2008). "Uncovering biological network function via graphlet degree signatures." Cancer Inform **6**: 257-273.

Milenković, T., V. Memisevic, A. Bonato and N. Pržulj (2011). "Dominating biological networks." Plos One **6**(8): e23016.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon (2002). "Network motifs: simple building blocks of complex networks." Science **298**(5594): 824-827.

Nakazawa, Y., K. Sasaki, N. Mitsutake, M. Matsuse, M. Shimada, T. Nardo, Y. Takahashi, K. Ohyama, K. Ito, H. Mishima, M. Nomura, A. Kinoshita, S. Ono, K. Takenaka, R. Masuyama, T. Kudo, H. Slor, A. Utani, S. Tateishi, S. Yamashita, M. Stefanini, A. R. Lehmann, K. Yoshiura and T. Ogi (2012). "Mutations in UVSSA cause UV-sensitive syndrome and impair RNA polymerase IIo processing in transcription-coupled nucleotide-excision repair." Nat Genet **44**(5): 586-592.

National Human Genome Research Institute (2003). "2003 Release: International Consortium Completes Human Genome Project." Retrieved 30/11/2014, from http://www.genome.gov/11006929.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder and M. J. Daly (2011). "Testing for an unusual distribution of rare variants." PLoS Genet **7**(3): e1001322.

Neale, B. M., Y. Kou, L. Liu, A. Ma'ayan, K. E. Samocha, A. Sabo, C. F. Lin, C. Stevens, L. S. Wang, V. Makarov, P. Polak, S. Yoon, J. Maguire, E. L. Crawford, N. G. Campbell, E. T. Geller, O. Valladares, C. Schafer, H. Liu, T. Zhao, G. Cai, J. Lihm, R. Dannenfelser, O. Jabado, Z. Peralta, U. Nagaswamy, D. Muzny, J. G. Reid, I. Newsham, Y. Wu, et al. (2012). "Patterns and rates of exonic de novo mutations in autism spectrum disorders." Nature **485**(7397): 242-245.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson and J. Shendure (2009). "Targeted capture and massively parallel sequencing of 12 human exomes." Nature **461**(7261): 272-276.

Ng, S. B., A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper, H. C. Mefford, C. Lee, E. H. Turner, J. D. Smith, M. J. Rieder, K. Yoshiura, N. Matsumoto, T. Ohta, N. Niikawa, D. A. Nickerson, M. J. Bamshad and J. Shendure (2010a). "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome." Nat Genet **42**(9): 790-793.

Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure and M. J. Bamshad (2010b). "Exome sequencing identifies the cause of a mendelian disorder." Nat Genet **42**(1): 30-35.

Ng, S. B., D. A. Nickerson, M. J. Bamshad and J. Shendure (2010c). "Massively parallel sequencing and rare disease." Hum Mol Genet **19**(R2): R119-124.

NHLBI Exome Sequencing Project (2014). "Exome Variant Server." Retrieved 07/07/2014, from http://evs.gs.washington.edu/EVS/.

Novarino, G., A. G. Fenstermaker, M. S. Zaki, M. Hofree, J. L. Silhavy, A. D. Heiberg, M. Abdellateef, B. Rosti, E. Scott, L. Mansour, A. Masri, H. Kayserili, J. Y. Al-Aama, G. M. H. Abdel-Salam, A. Karminejad, M. Kara, B. Kara, B. Bozorgmehri, T. Ben-Omran, F. Mojahedi, I. G. E. D. Mahmoud, N. Bouslam, A. Bouhouche, A. Benomar, S. Hanein, L. Raymond, S. Forlani, M. Mascaro, L. Selim, N. Shehata, et al. (2014). "Exome Sequencing Links Corticospinal Motor Neuron Disease to Common Neurodegenerative Disorders." Science **343**(6170): 506-511.

O'Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang and G. J. Lyon (2013). "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing." Genome Medicine **5**(3): 28.

O'Roak, B. J., P. Deriziotis, C. Lee, L. Vives, J. J. Schwartz, S. Girirajan, E. Karakoc, A. P. Mackenzie, S. B. Ng, C. Baker, M. J. Rieder, D. A. Nickerson, R. Bernier, S. E. Fisher, J. Shendure and E. E. Eichler (2011). "Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations." Nat Genet **43**(6): 585-589.

O'Roak, B. J., L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernot, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure and E. E. Eichler (2012). "Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations." Nature **485**(7397): 246-250.

O'Sullivan, M. and C. O'Morain (2006). "Nutrition in inflammatory bowel disease." Best Pract Res Clin Gastroenterol **20**(3): 561-573.

Obayashi, T., S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta and K. Kinoshita (2008). "COXPRESdb: a database of coexpressed gene networks in mammals." Nucleic Acids Res **36**(Database issue): D77-82.

Obayashi, T., Y. Okamura, S. Ito, S. Tadaka, I. N. Motoike and K. Kinoshita (2013). "COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals." Nucleic Acids Res **41**(Database issue): D1014-1020.

Ogura, Y., D. K. Bonen, N. Inohara, D. L. Nicolae, F. F. Chen, R. Ramos, H. Britton, T. Moran, R. Karaliuskas, R. H. Duerr, J. P. Achkar, S. R. Brant, T. M. Bayless, B. S. Kirschner, S. B. Hanauer, G. Nunez and J. H. Cho (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature **411**(6837): 603-606.

Olsen, C., K. Fleming, N. Prendergast, R. Rubio, F. Emmert-Streib, G. Bontempi, B. Haibe-Kains and J. Quackenbush (2014). "Inference and validation of predictive gene networks from biomedical literature and gene expression data." Genomics **103**(5-6): 329-336.

Oracle (2011). "Arrays (Java Platform SE 6)." Retrieved 20/10/2014, from http://docs.oracle.com/javase/6/docs/api/java/util/Arrays.html.

Orchard, S., M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, et al. (2014). "The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases." Nucleic Acids Res **42**(Database issue): D358-363.

Oti, M. and H. G. Brunner (2007). "The modular nature of genetic diseases." Clin Genet **71**(1): 1-11.

Pabinger, S., A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M. R. Speicher, J. Zschocke and Z. Trajanoski (2014). "A survey of tools for variant analysis of next-generation genome sequencing data." Brief Bioinform **15**(2): 256-278.

Parkes, M., J. C. Barrett, N. J. Prescott, M. Tremelling, C. A. Anderson, S. A. Fisher, R. G. Roberts, E. R. Nimmo, F. R. Cummings, D. Soars, H. Drummond, C. W. Lees, S. A. Khawaja, R. Bagnall, D. A. Burke, C. E. Todhunter, T. Ahmad, C. M. Onnie, W. McArdle, D. Strachan, G. Bethel, C. Bryan, C. M. Lewis, P. Deloukas, A. Forbes, J. Sanderson, D. P. Jewell, J. Satsangi, J. C. Mansfield, L. Cardon, et al. (2007). "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility." Nat Genet **39**(7): 830-832.

Parla, J. S., I. Iossifov, I. Grabill, M. S. Spector, M. Kramer and W. R. McCombie (2011). "A comparative analysis of exome capture." Genome Biol **12**(9): R97.

Peltonen, L., M. Perola, J. Naukkarinen and A. Palotie (2006). "Lessons from studying monogenic disease for common disease." Hum Mol Genet **15 Spec No 1**: R67-74.

Petrovski, S., Q. Wang, E. L. Heinzen, A. S. Allen and D. B. Goldstein (2013). "Genic intolerance to functional variation and the interpretation of personal genomes." PLoS Genet **9**(8): e1003709.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom and A. Siepel (2010). "Detection of nonneutral substitution rates on mammalian phylogenies." Genome Res **20**(1): 110-121.

Polvi, A., T. Linnankivi, T. Kivela, R. Herva, J. P. Keating, O. Makitie, D. Pareyson, L. Vainionpaa, J. Lahtinen, I. Hovatta, H. Pihko and A. E. Lehesjoki (2012). "Mutations in CTC1, encoding the CTS telomere maintenance complex component 1, cause cerebroretinal microangiopathy with calcifications and cysts." Am J Hum Genet **90**(3): 540-549.

Price, A. L., G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L. J. Wei and S. R. Sunyaev (2010). "Pooled association tests for rare variants in exon-resequencing studies." Am J Hum Genet **86**(6): 832-838.

Pruitt, K. D., J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, et al. (2009). "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes." Genome Res **19**(7): 1316-1323.

Pruitt, K. D., G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, M. R. Murphy, N. A. O'Leary, S. Pujar, B. Rajput, S. H. Rangwala, L. D. Riddick, A. Shkeda, H. Sun, P. Tamez, R. E. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. R. Maglott, T. D. Murphy and J. M. Ostell (2014). "RefSeq: an update on mammalian reference sequences." Nucleic Acids Res **42**(Database issue): D756-763.

Pržulj, N., D. G. Corneil and I. Jurisica (2004). "Modeling interactome: scale-free or geometric?" Bioinformatics **20**(18): 3508-3515.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, et al. (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.

R Development Core Team (2013). "R: A language and environment for statistical computing." Vienna, Austria: R Foundation for Statistical Computing.

Rabbani, B., N. Mahdieh, K. Hosomichi, H. Nakaoka and I. Inoue (2012). "Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders." J Hum Genet **57**(10): 621-632.

Rademakers, R., M. Baker, A. M. Nicholson, N. J. Rutherford, N. Finch, A. Soto-Ortolaza, J. Lash, C. Wider, A. Wojtas, M. DeJesus-Hernandez, J. Adamson, N. Kouri, C. Sundal, E. A. Shuster, J. Aasly, J. MacKenzie, S. Roeber, H. A. Kretzschmar, B. F. Boeve, D. S. Knopman, R. C. Petersen, N. J. Cairns, B. Ghetti, S. Spina, J. Garbern, A. C. Tselis, R. Uitti, P. Das, J. A. Van Gerpen, J. F. Meschia, et al. (2012). "Mutations in the colony stimulating factor 1 receptor (CSF1R) gene cause hereditary diffuse leukoencephalopathy with spheroids." Nat Genet **44**(2): 200-205.

Raman, K. (2010). "Construction and analysis of protein-protein interaction networks." Autom Exp **2**(1): 2.

Ramensky, V., P. Bork and S. Sunyaev (2002). "Human non-synonymous SNPs: server and survey." Nucleic Acids Res **30**(17): 3894-3900.

Raychaudhuri, S., R. M. Plenge, E. J. Rossin, A. C. Ng, S. M. Purcell, P. Sklar, E. M. Scolnick, R. J. Xavier, D. Altshuler and M. J. Daly (2009). "Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions." PLoS Genet **5**(6): e1000534.

Ren, W. and Q. Zhao (2011). "A note on 'Algorithms for connected set cover problem and fault-tolerant connected set cover problem'." Theoretical Computer Science **412**(45): 6451-6454.

Riepe, F. G. (2009). "Clinical and molecular features of type 1 pseudohypoaldosteronism." Horm Res **72**(1): 1-9.

Rigaut, G., A. Shevchenko, B. Rutz, M. Wilm, M. Mann and B. Seraphin (1999). "A generic protein purification method for protein complex characterization and proteome exploration." Nat Biotechnol **17**(10): 1030-1032.

Rivas, M. A., M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, T. Fennell, A. Kirby, A. Latiano, P. Goyette, T. Green, J. Halfvarson, T. Haritunians, J. M. Korn, F. Kuruvilla, C. Lagace, B. Neale, K. S. Lo, P. Schumm, L. Torkvist, M. C. Dubinsky, S. R. Brant, M. S. Silverberg, R. H. Duerr, D. Altshuler, S. Gabriel, et al. (2011). "Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease." Nat Genet **43**(11): 1066-1073.

Robinson, P., S. Kohler, A. Oellrich, K. Wang, C. Mungall, S. E. Lewis, N. Washington, S. Bauer, D. S. Seelow, P. Krawitz, C. Gilissen, M. Haendel and D. Smedley (2013). "Improved exome prioritization of disease genes through cross species phenotype comparison." Genome Res.

Robinson, P. N., P. Krawitz and S. Mundlos (2011). "Strategies for exome and genome sequence data analysis in disease-gene discovery projects." Clin Genet **80**(2): 127-132.

Rossin, E. J., K. Lage, S. Raychaudhuri, R. J. Xavier, D. Tatar, Y. Benita, C. Cotsapas and M. J. Daly (2011). "Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology." PLoS Genet **7**(1): e1001273.

Saeed, R. and C. M. Deane (2006). "Protein protein interactions, evolutionary rate, abundance and age." BMC Bioinformatics **7**: 128.

Samuels, M. E. (2010). "Saturation of the human phenome." Curr Genomics **11**(7): 482-499.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Sauna, Z. E. and C. Kimchi-Sarfaty (2011). "Understanding the contribution of synonymous mutations to human disease." Nat Rev Genet **12**(10): 683-691.

Schäffer, A. A. (2013). "Digenic inheritance in medical genetics." J Med Genet **50**(10): 641-652.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Schwarz, J. M., C. Rodelsperger, M. Schuelke and D. Seelow (2010). "MutationTaster evaluates disease-causing potential of sequence alterations." Nat Methods **7**(8): 575-576.

Sedgewick, R. (2003). Breadth-First Search. Algorithms in Java, Third Edition, Part 5: Graph Algorithms. Boston, Mass.; London, Addison-Wesley**:** 121-131.

Shaheen, R., E. Faqeih, A. Sunker, H. Morsy, T. Al-Sheddi, H. E. Shamseldin, N. Adly, M. Hashem and F. S. Alkuraya (2011). "Recessive mutations in DOCK6, encoding the guanidine nucleotide exchange factor DOCK6, lead to abnormal actin cytoskeleton organization and Adams-Oliver syndrome." Am J Hum Genet **89**(2): 328-333.

Shaheen, R., M. Aglan, K. Keppler-Noreuil, E. Faqeih, S. Ansari, K. Horton, A. Ashour, M. S. Zaki, F. Al-Zahrani, A. M. Cueto-Gonzalez, G. Abdel-Salam, S. Temtamy and F. S. Alkuraya (2013). "Mutations in EOGT confirm the genetic heterogeneity of autosomal-recessive Adams-Oliver syndrome." Am J Hum Genet **92**(4): 598-604.

Shannon, C. E. (1948). "The Bell System Technical Journal." A Mathematical Theory of Communication **27**(3): 379-423.

Shen-Orr, S. S., R. Milo, S. Mangan and U. Alon (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nat Genet **31**(1): 64-68.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin (2001). "dbSNP: the NCBI database of genetic variation." Nucleic Acids Res **29**(1): 308-311.

Sifrim, A., J. K. J. Van Houdt, L. C. Tranchevent, B. Nowakowska, R. Sakai, G. A. Pavlopoulos, K. Devriendt, J. R. Vermeesch, Y. Moreau and J. Aerts (2012). "Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease." Genome Medicine **4**.

Sifrim, A., D. Popovic, L. C. Tranchevent, A. Ardeshirdavani, R. Sakai, P. Konings, J. R. Vermeesch, J. Aerts, B. De Moor and Y. Moreau (2013). "eXtasy: variant prioritization by genomic data fusion." Nat Methods **10**(11): 1083-1084.

Simpson, M. A., M. D. Irving, E. Asilmaz, M. J. Gray, D. Dafou, F. V. Elmslie, S. Mansour, S. E. Holder, C. E. Brain, B. K. Burton, K. H. Kim, R. M. Pauli, S. Aftimos, H. Stewart, C. A. Kim, M. Holder-Espinasse, S. P. Robertson, W. M. Drake and R. C. Trembath (2011). "Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss." Nat Genet **43**(4): 303-305.

Simpson, M. A., C. Deshpande, D. Dafou, L. E. Vissers, W. J. Woollard, S. E. Holder, G. Gillessen-Kaesbach, R. Derks, S. M. White, R. Cohen-Snuijf, S. G. Kant, L. H. Hoefsloot, W. Reardon, H. G. Brunner, E. M. Bongers and R. C. Trembath (2012). "De novo mutations of the gene encoding the histone acetyltransferase KAT6B cause Genitopatellar syndrome." Am J Hum Genet **90**(2): 290-294.

Smoot, M. E., K. Ono, J. Ruscheinski, P. L. Wang and T. Ideker (2011). "Cytoscape 2.8: new features for data integration and network visualization." Bioinformatics **27**(3): 431-432.

Snape, K. M., D. Ruddy, M. Zenker, W. Wuyts, M. Whiteford, D. Johnson, W. Lam and R. C. Trembath (2009). "The spectra of clinical phenotypes in aplasia cutis congenita and terminal transverse limb defects." Am J Med Genet A **149A**(8): 1860-1881.

Southgate, L., R. D. Machado, K. M. Snape, M. Primeau, D. Dafou, D. M. Ruddy, P. A. Branney, M. Fisher, G. J. Lee, M. A. Simpson, Y. He, T. Y. Bradshaw, B. Blaumeiser, W. S. Winship, W. Reardon, E. R. Maher, D. R. FitzPatrick, W. Wuyts, M. Zenker, N. Lamarche-Vane and R. C. Trembath (2011). "Gain-of-function mutations of ARHGAP31, a Cdc42/Rac1 GTPase regulator, cause syndromic cutis aplasia and limb anomalies." Am J Hum Genet **88**(5): 574-585.

Spirin, V. and L. A. Mirny (2003). "Protein complexes and functional modules in molecular networks." Proc Natl Acad Sci U S A **100**(21): 12123-12128.

Staiger, C., S. Cadot, R. Kooter, M. Dittrich, T. Muller, G. W. Klau and L. F. Wessels (2012). "A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer." Plos One **7**(4): e34796.

Stelzer, G., I. Dalah, T. I. Stein, Y. Satanower, N. Rosen, N. Nativ, D. Oz-Levi, T. Olender, F. Belinky, I. Bahir, H. Krug, P. Perco, B. Mayer, E. Kolker, M. Safran and D. Lancet (2011). "In-silico human genomics with GeneCards." Hum Genomics **5**(6): 709-717.

Stenson, P. D., M. Mort, E. V. Ball, K. Shaw, A. Phillips and D. N. Cooper (2014). "The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine." Hum Genet **133**(1): 1-9.

Stittrich, A. B., A. Lehman, D. L. Bodian, J. Ashworth, Z. Zong, H. Li, P. Lam, A. Khromykh, R. K. Iyer, J. G. Vockley, R. Baveja, E. S. Silva, J. Dixon, E. L. Leon, B. D. Solomon, G. Glusman, J. E. Niederhuber, J. C. Roach and M. S. Patel (2014). "Mutations in NOTCH1 Cause Adams-Oliver Syndrome." Am J Hum Genet **95**(3): 275-284.

Stone, E. A. and A. Sidow (2005). "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity." Genome Res **15**(7): 978-986.

Stratton, M. R., P. J. Campbell and P. A. Futreal (2009). "The cancer genome." Nature **458**(7239): 719-724.

Stuart, J. M., E. Segal, D. Koller and S. K. Kim (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-255.

Su, S. Y., J. Kasberger, S. Baranzini, W. Byerley, W. Liao, J. Oksenberg, E. Sherr and E. Jorgenson (2012). "Detection of identity by descent using next-generation whole genome sequencing data." BMC Bioinformatics **13**: 121.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-15550.

Sunyaev, S. R., F. Eisenhaber, I. V. Rodchenkov, B. Eisenhaber, V. G. Tumanyan and E. N. Kuznetsov (1999). "PSIC: profile extraction from sequence alignments with position-specific counts of independent observations." Protein Eng **12**(5): 387-394.

Tatton-Brown, K., S. Seal, E. Ruark, J. Harmer, E. Ramsay, S. Del Vecchio Duarte, A. Zachariou, S. Hanks, E. O'Brien, L. Aksglaede, D. Baralle, T. Dabir, B. Gener, D. Goudie, T. Homfray, A. Kumar, D. T. Pilz, A. Selicorni, I. K. Temple, L. Van Maldergem, N. Yachelevich, R. van Montfort and N. Rahman (2014). "Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability." Nat Genet **46**(4): 385-388.

Torres, E. M., B. R. Williams and A. Amon (2008). "Aneuploidy: cells losing their balance." Genetics **179**(2): 737-746.

Toydemir, R. M., A. Rutherford, F. G. Whitby, L. B. Jorde, J. C. Carey and M. J. Bamshad (2006). "Mutations in embryonic myosin heavy chain (MYH3) cause Freeman-Sheldon syndrome and Sheldon-Hall syndrome." Nat Genet **38**(5): 561-565.

Travassos, L. H., L. A. Carneiro, M. Ramjeet, S. Hussey, Y. G. Kim, J. G. Magalhaes, L. Yuan, F. Soares, E. Chea, L. Le Bourhis, I. G. Boneca, A. Allaoui, N. L. Jones, G. Nunez, S. E. Girardin and D. J. Philpott (2010). "Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry." Nat Immunol **11**(1): 55-62.

Turner, D. J., T. M. Keane, I. Sudbery and D. J. Adams (2009). "Next-generation sequencing of vertebrate experimental organisms." Mamm Genome **20**(6): 327-338.

Uhlig, H. H., T. Schwerd, S. Koletzko, N. Shah, J. Kammermeier, A. Elkadri, J. Ouahed, D. C. Wilson, S. P. Travis, D. Turner, C. Klein, S. B. Snapper and A. M. Muise (2014). "The Diagnostic Approach to Monogenic Very Early Onset Inflammatory Bowel Disease." Gastroenterology.

UK10K Consortium (2011). "UK10K - UK10K Goals." Retrieved 07/07/2014, from http://www.uk10k.org/goals.html.

Ulitsky, I. and R. Shamir (2007). "Identification of functional modules using network topology and high-throughput data." BMC Syst Biol **1**: 8.

Ulitsky, I., A. Krishnamurthy, R. M. Karp and R. Shamir (2010). "DEGAS: de novo discovery of dysregulated pathways in human diseases." Plos One **5**(10): e13367.

van Bodegraven, A. A., C. R. Curley, K. A. Hunt, A. J. Monsuur, R. K. Linskens, C. M. Onnie, J. B. Crusius, V. Annese, A. Latiano, M. S. Silverberg, A. Bitton, S. A. Fisher, A. H. Steinhart, A. Forbes, J. Sanderson, N. J. Prescott, D. P. Strachan, R. J. Playford, C. G. Mathew, C. Wijmenga, M. J. Daly, J. D. Rioux and D. A. van Heel (2006). "Genetic variation in myosin IXB is associated with ulcerative colitis." Gastroenterology **131**(6): 1768-1774.

van Driel, M. A., J. Bruggeman, G. Vriend, H. G. Brunner and J. A. Leunissen (2006). "A text-mining analysis of the human phenome." Eur J Hum Genet **14**(5): 535-542.

van Kasteren, S. I., H. Overkleeft, H. Ovaa and J. Neefjes (2014). "Chemical biology of antigen presentation by MHC molecules." Curr Opin Immunol **26**: 21-31.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Vidal, M., M. E. Cusick and A. L. Barabasi (2011). "Interactome networks and human disease." Cell **144**(6): 986-998.

Walsh, T., H. Shahin, T. Elkan-Miller, M. K. Lee, A. M. Thornton, W. Roeb, A. Abu Rayyan, S. Loulus, K. B. Avraham, M. C. King and M. Kanaan (2010). "Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82." Am J Hum Genet **87**(1): 90-94.

Wang, B., W. Yang, W. Wen, J. Sun, B. Su, B. Liu, D. Ma, D. Lv, Y. Wen, T. Qu, M. Chen, M. Sun, Y. Shen and X. Zhang (2010a). "Gamma-secretase gene mutations in familial acne inversa." Science **330**(6007): 1065.

Wang, J., D. Duncan, Z. Shi and B. Zhang (2013a). "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013." Nucleic Acids Res **41**(Web Server issue): W77-83.

Wang, J. L., X. Yang, K. Xia, Z. M. Hu, L. Weng, X. Jin, H. Jiang, P. Zhang, L. Shen, J. F. Guo, N. Li, Y. R. Li, L. F. Lei, J. Zhou, J. Du, Y. F. Zhou, Q. Pan, J. Wang, R. Q. Li and B. S. Tang (2010b). "TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing." Brain **133**(Pt 12): 3510-3518.

Wang, K., M. Li and H. Hakonarson (2010c). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Res **38**(16): e164.

Wang, L., X. Wang, A. P. Arkin and M. S. Samoilov (2013b). "Inference of gene regulatory networks from genome-wide knockout fitness data." Bioinformatics **29**(3): 338-346.

Wang, X., X. Wei, B. Thijssen, J. Das, S. M. Lipkin and H. Yu (2012). "Three-dimensional reconstruction of protein networks provides insight into human genetic disease." Nat Biotechnol **30**(2): 159-164.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Warde-Farley, D., S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader and Q. Morris (2010). "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function." Nucleic Acids Res **38**(Web Server issue): W214-220.

Weatherall, D. J. (2001). "Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias." Nat Rev Genet **2**(4): 245-255.

Weinshilboum, R. (2003). "Inheritance and drug response." N Engl J Med **348**(6): 529-537.

Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.

Winand, R., K. Hens, W. Dondorp, G. de Wert, Y. Moreau, J. R. Vermeesch, I. Liebaers and J. Aerts (2014). "In vitro screening of embryos by whole-genome sequencing: now, in the future or never?" Hum Reprod.

Windheim, M., C. Lang, M. Peggie, L. A. Plater and P. Cohen (2007). "Molecular mechanisms involved in the regulation of cytokine production by muramyl dipeptide." Biochem J **404**(2): 179-190.

Wishart, D. S., T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner and A. Scalbert (2013). "HMDB 3.0--The Human Metabolome Database in 2013." Nucleic Acids Res **41**(Database issue): D801-807.

Wu, G. and D. Zhi (2013). "Pathway-Based Approaches for Sequencing-Based Genome-Wide Association Studies." Genet Epidemiol.

Wu, J., Y. Li and R. Jiang (2014). "Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies." PLoS Genet **10**(3): e1004237.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke and X. Lin (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." Am J Hum Genet **89**(1): 82-93.

Xu, K., I. Bezakova, L. Bunimovich and S. V. Yi (2011). "Path lengths in protein-protein interaction networks and biological complexity." Proteomics **11**(10): 1857-1867.

Yamada, T. and P. Bork (2009). "Evolution of biomolecular networks: lessons from metabolic and protein interactions." Nat Rev Mol Cell Biol **10**(11): 791-803.

Yamashita, A. S., M. V. Geraldo, C. S. Fuziwara, M. A. Kulcsar, C. U. Friguglietti, R. B. da Costa, G. S. Baia and E. T. Kimura (2013). "Notch pathway is activated by MAPK signaling and influences papillary thyroid cancer proliferation." Transl Oncol **6**(2): 197-205.

York, B. and B. W. O'Malley (2010). "Steroid receptor coactivator (SRC) family: masters of systems biology." J Biol Chem **285**(50): 38743-38750.

Yu, H., L. Tardivo, S. Tam, E. Weiner, F. Gebreab, C. Fan, N. Svrzikapa, T. Hirozane-Kishikawa, E. Rietman, X. Yang, J. Sahalie, K. Salehi-Ashtiani, T. Hao, M. E. Cusick, D. E. Hill, F. P. Roth, P. Braun and M. Vidal (2011). "Next-generation sequencing to generate interactome datasets." Nat Methods **8**(6): 478-480.

Zhang, Y. Z. and Y. Y. Li (2014). "Inflammatory bowel disease: pathogenesis." World J Gastroenterol **20**(1): 91-99.

Zhang, Z., X. F. Gao and W. L. Wu (2009). "Algorithms for connected set cover problem and fault-tolerant connected set cover problem." Theoretical Computer Science **410**(8-10): 812-817.

Zhong, Q., N. Simonis, Q. R. Li, B. Charloteaux, F. Heuze, N. Klitgord, S. Tam, H. Yu, K. Venkatesan, D. Mou, V. Swearingen, M. A. Yildirim, H. Yan, A. Dricot, D. Szeto, C. Lin, T. Hao, C. Fan, S. Milstein, D. Dupuy, R. Brasseur, D. E. Hill, M. E. Cusick and M. Vidal (2009). "Edgetic perturbation models of human inherited disorders." Mol Syst Biol **5**: 321.

Zhuang, Z., A. Gusev, J. Cho and I. Pe'er (2012). "Detecting identity by descent and homozygosity mapping in whole-exome sequencing data." Plos One **7**(10): e47618.

Zlotogora, J. (2003). "Penetrance and expressivity in the molecular age." Genet Med **5**(5): 347-352.

Appendix A: OMIM Disease Subnetworks

The following table lists all disease subnetworks identified in the PINA and PINAmin2 networks. Disease subnetworks are connected sets of genes causing the same monogenic disease. See chapter 3 for full description. Note that generalised disease terms have had disease "type" or "group" designations removed

Network	**Generalised disease term**	**Causal genes connected in network**
PINA	3-M syndrome	*CUL7, OBSL1*
PINAmin2	3-M syndrome	*CUL7, OBSL1*
PINA	46XY sex reversal	*MAP3K1, SRY*
PINA	Acne inversa familial	*PSEN1, NCSTN, PSENEN*
PINAmin2	Acne inversa familial	*PSEN1, NCSTN, PSENEN*
PINA	Afibrinogenemia congenital	*FGB, FGA*
PINAmin2	Afibrinogenemia congenital	*FGA, FGB*
PINA	Agammaglobulinemia	*CD79A, CD79B, BLNK, IGHM*
PINAmin2	Agammaglobulinemia	*CD79A, BLNK, CD79B*
PINA	Albinism oculocutaneous type	*TYR, TYRP1*
PINAmin2	Albinism oculocutaneous type	*TYR, TYRP1*
PINA	Arrhythmogenic right ventricular dysplasia	*JUP, DSP, DSC2, DSC3, DSG2, PKP2*
PINAmin2	Arrhythmogenic right ventricular dysplasia	*JUP, DSP*
PINA	Arthrogryposis renal dysfunction and cholestasis	*VPS33B, VIPAS39*
PINAmin2	Arthrogryposis renal dysfunction and cholestasis	*VPS33B, VIPAS39*
PINA	Atrial fibrillation familial	*KCNQ1, KCNE2*
PINA	Axenfeld-Rieger syndrome type	*FOXC1, PITX2*
PINA	Baraitser-Winter syndrome	*ACTB, ACTG1*
PINAmin2	Baraitser-Winter syndrome	*ACTB, ACTG1*
PINA	Bardet-Biedl syndrome	*ARL6, BBS1, BBS4, BBS12, BBS7, BBS9, BBS10, BBS2, MKKS, BBS5, TTC8*
PINAmin2	Bardet-Biedl syndrome	*BBS1, BBS4, BBS7, BBS9, BBS12, BBS2, ARL6, MKKS, BBS5, TTC8*
PINA	Bare lymphocyte syndrome type	*TAP1, TAPBP*
PINAmin2	Bare lymphocyte syndrome type	*TAPBP, TAP1*
PINA	Bare lymphocyte syndrome type complementation group	*RFX5, RFXAP*
PINA	Basal cell carcinoma somatic	*PTCH1, SMO, PTCH2*
PINA	Bernard-Soulier syndrome type	*GP9, GP1BB*
PINAmin2	Bernard-Soulier syndrome type	*GP1BB, GP9*

Network	Generalised disease term	Causal genes connected in network
PINA	Brachydactyly type	*BMPR1B, BMP2, GDF5, NOG*
PINAmin2	Brachydactyly type	*BMP2, BMPR1B, GDF5, NOG*
PINA	Bradyopsia	*RGS9, RGS9BP*
PINA	Breast cancer	*TP53, ESR1, PPM1D*
PINA	Bronchiectasis with or without elevated sweat chloride	*SCNN1B, SCNN1A, SCNN1G*
PINA	Brugada syndrome	*CACNA1C, CACNB2*
PINA	C1q deficiency	*C1QA, C1QB, C1QC*
PINAmin2	C1q deficiency	*C1QA, C1QB, C1QC*
PINA	C8 deficiency type	*C8A, C8B*
PINA	Cardiofaciocutaneous syndrome	*BRAF, MAP2K1, MAP2K2*
PINAmin2	Cardiofaciocutaneous syndrome	*BRAF, MAP2K2, MAP2K1*
PINA	Cardiomyopathy familial hypertrophic	*TNNI3, TPM1, TNNT2, TNNC1*
PINA	Cardiomyopathy familial hypertrophic	*MYBPC3, TTN*
PINAmin2	Cardiomyopathy familial hypertrophic	*TNNI3, TNNT2, TNNC1*
PINA	Cataract Coppock-like	*CRYBB2, CRYGC*
PINA	Cerebrooculofacioskeletal syndrome	*ERCC2, ERCC6, ERCC5*
PINAmin2	Cerebrooculofacioskeletal syndrome	*ERCC5, ERCC6*
PINA	Charcot-Marie-Tooth disease axonal type	*HSPB1, HSPB8*
PINAmin2	Charcot-Marie-Tooth disease axonal type	*HSPB1, HSPB8*
PINA	Charcot-Marie-Tooth disease type	*NEFL, MTMR2*
PINA	Charcot-Marie-Tooth disease type	*PMP22, MPZ*
PINA	Choriodal dystrophy central areolar	*PRPH2, PRPH*
PINA	Cirrhosis cryptogenic	*KRT18, KRT8*
PINAmin2	Cirrhosis cryptogenic	*KRT8, KRT18*
PINA	Cockayne syndrome type	*ERCC6, ERCC8*
PINAmin2	Cockayne syndrome type	*ERCC8, ERCC6*
PINA	Colorectal cancer hereditary nonpolyposis type	*MLH1, MSH2, PMS2, MSH6*
PINAmin2	Colorectal cancer hereditary nonpolyposis type	*MLH1, MSH2, MSH6, PMS2*
PINA	Colorectal cancer somatic	*EP300, APC, CTNNB1, AKT1, BRAF, BUB1B, DLC1, AXIN2*
PINA	Colorectal cancer somatic	*NRAS, PIK3CA*
PINAmin2	Colorectal cancer somatic	*AKT1, APC, CTNNB1, EP300, AXIN2*
PINA	Combined cellular and humoral immune defects with granulomas	*RAG1, RAG2*
PINA	Cone-rod dystrophy	*GUCA1A, GUCY2D*
PINAmin2	Cone-rod dystrophy	*GUCA1A, GUCY2D*
PINA	Congenital disorder of glycosylation type	*DPM1, DPM2, DPM3*
PINA	Congenital disorder of glycosylation type	*COG1, COG4, COG7, COG5, COG6*
PINAmin2	Congenital disorder of glycosylation type	*COG1, COG4, COG7, COG6, COG5*
PINAmin2	Congenital disorder of glycosylation type	*DPM1, DPM3*

Network	Generalised disease term	Causal genes connected in network
PINA	Cornelia de Lange syndrome	*RAD21, SMC3, NIPBL*
PINAmin2	Cornelia de Lange syndrome	*RAD21, SMC3*
PINA	Cowden syndrome	*AKT1, PTEN*
PINA	Cutis laxa autosomal recessive type	*EFEMP2, FBLN5*
PINA	Deafness autosomal dominant	*SIX1, EYA4*
PINA	Deafness autosomal recessive	*MYO7A, CDH23, DFNB31, MYO15A*
PINA	Dejerine-Sottas disease	*PMP22, MPZ*
PINA	Dementia Lewy body	*SNCA, SNCB*
PINA	Diabetes mellitus permanent neonatal	*INS, GCK*
PINA	Diabetes mellitus transient neonatal	*ABCC8, KCNJ11*
PINA	Diamond-Blackfan anemia	*RPL5, RPL11*
PINAmin2	Diamond-Blackfan anemia	*RPL5, RPL11*
PINA	Dysfibrinogenemia type	*FGB, FGG*
PINAmin2	Dysfibrinogenemia type	*FGB, FGG*
PINA	Ehlers-Danlos syndrome type	*COL1A1, COL1A2*
PINA	Ehlers-Danlos syndrome type	*COL5A2, COL5A1*
PINAmin2	Ehlers-Danlos syndrome type	*COL1A1, COL1A2*
PINA	Epidermolysis bullosa simplex type	*KRT5, KRT14*
PINAmin2	Epidermolysis bullosa simplex type	*KRT14, KRT5*
PINA	Epidermolysis bullosa junctional type	*ITGB4, COL17A1*
PINA	Epidermolysis bullosa junctional type	*LAMB3, LAMA3, LAMC2*
PINAmin2	Epidermolysis bullosa junctional type	*LAMB3, LAMA3, LAMC2*
PINA	Epilepsy nocturnal frontal lobe	*CHRNB2, CHRNA4*
PINAmin2	Epilepsy nocturnal frontal lobe	*CHRNB2, CHRNA4*
PINA	Epilepsy progressive myoclonic 2B (Lafora)	*EPM2A, NHLRC1*
PINAmin2	Epilepsy progressive myoclonic 2B (Lafora)	*EPM2A, NHLRC1*
PINA	Epiphyseal dysplasia multiple	*COL9A1, COMP, MATN3*
PINA	Episodic ataxia type	*CACNA1A, CACNB4*
PINA	Erythrocytosis familial	*VHL, EPAS1, EGLN1*
PINAmin2	Erythrocytosis familial	*VHL, EPAS1, EGLN1*
PINA	Exostoses multiple type	*EXT1, EXT2*
PINA	Fanconi anemia complementation group	*FANCA, FANCG, FANCM, BRCA2, PALB2, FANCC, FANCB, FANCI, FANCD2, FANCE, FANCF, BRIP1*
PINAmin2	Fanconi anemia complementation group	*FANCG, BRCA2, FANCA, FANCM, FANCC, FANCD2, FANCE, PALB2, FANCF, FANCB, FANCI*
PINA	Foveomacular dystrophy adult-onset with choroidal neovascularization	*PRPH2, PRPH*
PINA	Frontonasal dysplasia	*ALX1, ALX4*
PINA	Glanzmann thrombasthenia	*ITGB3, ITGA2B*
PINAmin2	Glanzmann thrombasthenia	*ITGA2B, ITGB3*
PINA	Griscelli syndrome type	*RAB27A, MLPH, MYO5A*
PINAmin2	Griscelli syndrome type	*MLPH, RAB27A, MYO5A*
PINA	Hemangioma capillary infantile somatic	*KDR, FLT4*

Network	Generalised disease term	Causal genes connected in network
PINAmin2	Hemangioma capillary infantile somatic	*KDR, FLT4*
PINA	Hemochromatosis type	*HAMP, SLC40A1*
PINA	Hemophagocytic lymphohistiocytosis familial	*STX11, FHL5*
PINA	Hepatocellular carcinoma somatic	*CTNNB1, AXIN1, CASP8*
PINAmin2	Hepatocellular carcinoma somatic	*AXIN1, CTNNB1*
PINA	Hermansky-Pudlak syndrome	*DTNBP1, BLOC1S3*
PINA	Hermansky-Pudlak syndrome	*HPS1, HPS4*
PINA	Hermansky-Pudlak syndrome	*HPS6, HPS5*
PINAmin2	Hermansky-Pudlak syndrome	*BLOC1S3, DTNBP1*
PINAmin2	Hermansky-Pudlak syndrome	*HPS6, HPS5*
PINA	Hyperinsulinemic hypoglycemia familial	*ABCC8, KCNJ11*
PINA	Hypogonadotropic hypogonadism with or without anosmia	*FGFR1, FGF8*
PINA	Hypogonadotropic hypogonadism with or without anosmia	*GNRH1, GNRHR*
PINA	Hypogonadotropic hypogonadism with or without anosmia	*TACR3, TAC3*
PINA	Hypogonadotropic hypogonadism with or without anosmia	*KISS1, KISS1R*
PINA	Immune dysfunction with T-cell inactivation due to calcium entry defect	*ORAI1, STIM1*
PINAmin2	Immune dysfunction with T-cell inactivation due to calcium entry defect	*ORAI1, STIM1*
PINA	Immunodeficiency common variable	*CD19, CR2, CD81*
PINAmin2	Immunodeficiency common variable	*CD19, CR2, CD81*
PINA	Iridogoniodysgenesis type	*FOXC1, PITX2*
PINA	Kabuki syndrome	*KDM6A, MLL2*
PINAmin2	Kabuki syndrome	*KDM6A, MLL2*
PINA	LADD syndrome	*FGFR2, FGF10*
PINAmin2	LADD syndrome	*FGF10, FGFR2*
PINA	LEOPARD syndrome	*RAF1, BRAF*
PINAmin2	LEOPARD syndrome	*RAF1, BRAF*
PINA	Leigh syndrome due to mitochondrial complex I deficiency	*NDUFS3, NDUFA9*
PINAmin2	Leigh syndrome due to mitochondrial complex I deficiency	*NDUFS3, NDUFA9*
PINA	Leukemia acute myeloid	*CEBPA, RUNX1*
PINA	Leukemia acute promyelocytic type	*ZBTB16, PML*
PINA	Leukoencephalopathy with vanishing white matter	*EIF2B1, EIF2B2, EIF2B5, EIF2B3*
PINAmin2	Leukoencephalopathy with vanishing white matter	*EIF2B5, EIF2B1*
PINA	Li-Fraumeni syndrome	*CHEK2, TP53*
PINAmin2	Li-Fraumeni syndrome	*TP53, CHEK2*
PINA	Liddle syndrome	*SCNN1B, SCNN1G*
PINA	Lissencephaly	*TUBA1A, PAFAH1B1*
PINA	Loeys-Dietz syndrome type	*SMAD3, TGFBR1, TGFBR2, TGFB2*

Network	Generalised disease term	Causal genes connected in network
PINAmin2	Loeys-Dietz syndrome type	*TGFBR1, SMAD3, TGFBR2, TGFB2*
PINA	MODY type	*HNF4A, HNF1A*
PINA	Macular dystrophy	*PRPH2, PRPH*
PINA	Macular dystrophy patterned	*PRPH2, PRPH*
PINA	Macular dystrophy vitelliform	*PRPH2, PRPH*
PINA	Maple syrup urine disease type	*BCKDHA, BCKDHB*
PINAmin2	Maple syrup urine disease type	*BCKDHA, BCKDHB*
PINA	Meckel syndrome	*TMEM67, MKS1*
PINA	Meier-Gorlin syndrome	*CDC6, CDT1, ORC4, ORC1, ORC6*
PINAmin2	Meier-Gorlin syndrome	*CDC6, ORC1, ORC4, CDT1, ORC6*
PINA	Mental retardation X-linked	*GDI1, FTSJ1*
PINA	Mental retardation autosomal dominant	*SMARCA4, ARID1A, CTNNB1, GRIN1, GRIN2B, ARID1B, SMARCB1, CDH15, SYNGAP1, CDH3, CACNG2*
PINAmin2	Mental retardation autosomal dominant	*SMARCA4, ARID1A, SMARCB1, ARID1B*
PINAmin2	Mental retardation autosomal dominant	*CTNNB1, CDH3*
PINAmin2	Mental retardation autosomal dominant	*GRIN2B, GRIN1*
PINA	Methemoglobinemia type	*CYB5A, CYB5R3*
PINA	Microcephaly primary autosomal recessive	*CEP135, CENPJ*
PINA	Mismatch repair cancer syndrome	*MLH1, MSH2, PMS2, MSH6*
PINAmin2	Mismatch repair cancer syndrome	*MLH1, MSH2, MSH6, PMS2*
PINA	Mitochondrial complex I deficiency	*NDUFS3, NDUFS2*
PINAmin2	Mitochondrial complex I deficiency	*NDUFS3, NDUFS2*
PINA	Muir-Torre syndrome	*MLH1, MSH2*
PINAmin2	Muir-Torre syndrome	*MLH1, MSH2*
PINA	Multiple pterygium syndrome type	*CHRND, CHRNA1, CHRNG*
PINAmin2	Multiple pterygium syndrome type	*CHRNA1, CHRND*
PINA	Muscular dystrophy limb-girdle type	*DYSF, TCAP, DNAJB6, CAPN3, TTN, CAV3, SGCG, SGCA, SGCB, SGCD*
PINAmin2	Muscular dystrophy limb-girdle type	*TCAP, CAPN3, TTN*
PINAmin2	Muscular dystrophy limb-girdle type	*SGCG, SGCB, SGCA, SGCD*
PINA	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies) type	*POMT2, POMT1*
PINA	Muscular dystrophy-dystroglycanopathy (congenital with mental retardation) type	*POMT2, POMT1*
PINA	Muscular dystrophy-dystroglycanopathy (limb-girdle) type	*POMT2, POMT1*
PINA	Myasthenic syndrome congenital associated with acetylcholine receptor deficiency	*CHRNB1, MUSK, RAPSN*
PINA	Myasthenic syndrome fast-channel congenital	*CHRND, CHRNA1, CHRNE*

Network	Generalised disease term	Causal genes connected in network
PINAmin2	Myasthenic syndrome fast-channel congenital	*CHRNA1, CHRND*
PINA	Myasthenic syndrome slow-channel congenital	*CHRND, CHRNA1, CHRNE*
PINAmin2	Myasthenic syndrome slow-channel congenital	*CHRNA1, CHRND*
PINA	Mycobacterial infection atypical familial disseminated	*STAT1, IFNGR1*
PINAmin2	Mycobacterial infection atypical familial disseminated	*STAT1, IFNGR1*
PINA	Myopathy congenital with fiber-type	*ACTA1, TPM3*
PINA	Nasu-Hakola disease	*TYROBP, TREM2*
PINA	Nephrolithiasis/osteoporosis hypophosphatemic	*SLC9A3R1, SLC34A1*
PINA	Nephrotic syndrome type	*NPHS1, NPHS2*
PINA	Neuropathy distal hereditary motor type	*HSPB1, HSPB8*
PINAmin2	Neuropathy distal hereditary motor type	*HSPB1, HSPB8*
PINA	Neuropathy hereditary sensory and autonomic type	*SPTLC1, SPTLC2*
PINA	Noonan syndrome	*SOS1, PTPN11*
PINA	Noonan syndrome	*NRAS, KRAS, RAF1, BRAF*
PINAmin2	Noonan syndrome	*SOS1, PTPN11*
PINAmin2	Noonan syndrome	*RAF1, BRAF, KRAS, NRAS*
PINA	Omenn syndrome	*RAG1, RAG2*
PINA	Orofacial cleft	*SUMO1, MSX1, TP63*
PINA	Osteogenesis imperfecta type	*COL1A1, BMP1, COL1A2*
PINAmin2	Osteogenesis imperfecta type	*COL1A1, COL1A2*
PINA	Osteopetrosis autosomal recessive	*TNFSF11, TNFRSF11A*
PINA	Osteopetrosis autosomal recessive	*CLCN7, OSTM1*
PINA	Osteosarcoma somatic	*CHEK2, RB1*
PINA	Ovarian cancer somatic	*ERBB2, CTNNB1*
PINA	Ovarioleukodystrophy	*EIF2B2, EIF2B5, EIF2B4*
PINA	Pachyonychia congenita type	*KRT17, KRT6A*
PINA	Pancreatic cancer	*TP53, BRCA2*
PINAmin2	Pancreatic cancer	*TP53, BRCA2*
PINA	Paragangliomas	*SDHA, SDHB*
PINAmin2	Paragangliomas	*SDHB, SDHA*
PINA	Parkinson disease	*HTRA2, EIF4G1*
PINA	Parkinson disease	*SNCA, LRRK2*
PINA	Persistent Mullerian duct syndrome type	*AMH, AMHR2*
PINA	Pick disease	*MAPT, PSEN1*
PINA	Pituitary hormone deficiency combined	*PROP1, HESX1*
PINA	Pontocerebellar hypoplasia type	*TSEN2, TSEN54, TSEN34*
PINA	Porencephaly	*COL4A1, COL4A2*
PINAmin2	Porencephaly	*COL4A1, COL4A2*
PINA	Propionicacidemia	*PCCB, PCCA*
PINA	Pseudohypoaldosteronism type	*SCNN1B, SCNN1A, SCNN1G*
PINA	Renal tubular dysgenesis	*AGT, AGTR1*

Network	Generalised disease term	Causal genes connected in network
PINA	Retinitis pigmentosa	*ABCA4, CNGB1, PRPH2, PRPH*
PINA	Retinitis pigmentosa digenic	*PRPH2, PRPH, ROM1*
PINA	Retinitis punctata albescens	*PRPH2, PRPH*
PINA	Roussy-Levy syndrome	*PMP22, MPZ*
PINA	Severe combined immunodeficiency B cell-negative	*RAG1, RAG2*
PINA	Severe combined immunodeficiency T cell-negative B-cell/natural killer-cell positive	*CD3D, CD3E*
PINAmin2	Severe combined immunodeficiency T cell-negative B-cell/natural killer-cell positive	*CD3D, CD3E*
PINA	Sitosterolemia	*ABCG8, ABCG5*
PINAmin2	Sitosterolemia	*ABCG8, ABCG5*
PINA	Spastic paraplegia autosomal recessive	*AP4E1, AP4M1, AP4S1, AP4B1*
PINAmin2	Spastic paraplegia autosomal recessive	*AP4E1, AP4B1*
PINA	Spherocytosis type	*SPTA1, SPTB*
PINA	Spherocytosis type	*SLC4A1, ANK1*
PINAmin2	Spherocytosis type	*SPTA1, SPTB*
PINA	Spinocerebellar ataxia	*ATXN1, ATXN2*
PINA	Stickler syndrome type	*COL2A1, COL9A2*
PINA	Surfactant metabolism dysfunction pulmonary	*CSF2RB, CSF2RA*
PINAmin2	Surfactant metabolism dysfunction pulmonary	*CSF2RA, CSF2RB*
PINA	Telangiectasia hereditary hemorrhagic type	*ENG, ACVRL1*
PINA	Thrombocythemia	*JAK2, MPL, THPO*
PINAmin2	Thrombocythemia	*THPO, MPL*
PINA	Thrombophilia dysfibrinogenemic	*FGB, FGG*
PINAmin2	Thrombophilia dysfibrinogenemic	*FGB, FGG*
PINA	Thyroid carcinoma papillary	*TRIM24, TRIM33*
PINA	Treacher Collins syndrome	*POLR1C, POLR1D*
PINAmin2	Treacher Collins syndrome	*POLR1C, POLR1D*
PINA	Trichothiodystrophy	*ERCC2, ERCC3*
PINAmin2	Trichothiodystrophy	*ERCC3, ERCC2*
PINA	UV-sensitive syndrome	*ERCC6, ERCC8*
PINAmin2	UV-sensitive syndrome	*ERCC8, ERCC6*
PINA	Usher syndrome type	*MYO7A, CDH23, DFNB31, USH1C, USH1G*
PINAmin2	Usher syndrome type	*CDH23, USH1C*
PINA	Ventricular septal defect	*GATA4, NKX2-5*
PINAmin2	Ventricular septal defect	*GATA4, NKX2-5*
PINA	Waardenburg syndrome type	*EDNRB, EDN3*
PINA	Waardenburg syndrome type	*SOX10, PAX3, MITF*
PINAmin2	Waardenburg syndrome type	*PAX3, SOX10*
PINA	Warburg micro syndrome	*RAB3GAP2, RAB3GAP1*
PINA	Xeroderma pigmentosum group	*ERCC2, ERCC3, XPA, XPC, ERCC5, ERCC4*
PINAmin2	Xeroderma pigmentosum group	*ERCC3, ERCC2, ERCC5*

Appendix B: Example of a Network in which Both of BioGranat-IG's Heuristic Searches Fail

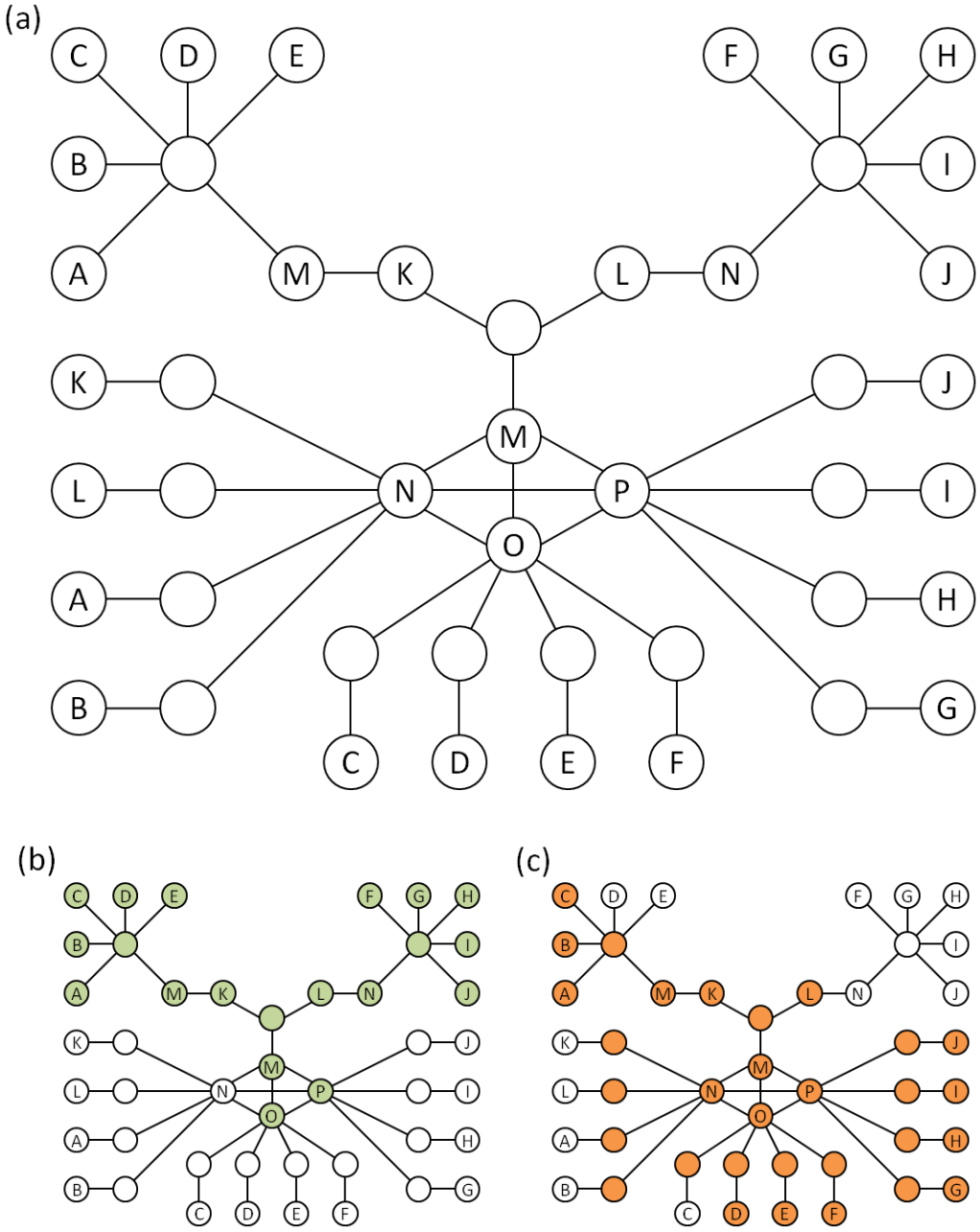


Figure B.1 – Example of a network in which all of BioGranat-IG's search algorithms fail to find the optimal subnetwork
See next page for full figure legend.

Figure B.1 – Example of a network in which all of BioGranat-IG's search algorithms fail to find the optimal subnetwork (previous page)
(a) An example of a network for which both the minimum distance search and multi-minimum distance search would fail to find the smallest subnetwork containing all individuals A-P. Nodes are labelled with the individuals attached to them. (b) The true optimal subnetwork comprises 20 nodes in the top section of the graph, shown in green. (c) However, nodes from the bottom section will be incorporated into any subnetwork found by the minimum distance and multi-minimum distance searches, regardless of which node the search starts from. The smallest subnetworks found by BioGranat-IG contain 31 nodes. There were 50 alternative such subnetworks, one of which is shown here in orange.

Appendix C: Supporting Publication

Nick Dand, Frauke Sprengel, Volker Ahlers and Thomas Schlitt, *BioGranat-IG: a network analysis tool to suggest mechanisms of genetic heterogeneity from exome-sequencing data*, Bioinformatics 29(6) (2013), pp. 733-41.

This publication is available online at:
http://bioinformatics.oxfordjournals.org/content/29/6/733.full.pdf